

Metastable energy strata in numerical discretizations of weakly nonlinear wave equations

Ludwig Gauckler¹ Daniel Weiss²

Version of 9 December 2016

Abstract

The quadratic nonlinear wave equation on a one-dimensional torus with small initial values located in a single Fourier mode is considered. In this situation, the formation of metastable energy strata has recently been described and their long-time stability has been shown. The topic of the present paper is the correct reproduction of these metastable energy strata by a numerical method. For symplectic trigonometric integrators applied to the equation, it is shown that these energy strata are reproduced even on long time intervals in a qualitatively correct way.

Mathematics Subject Classification (2010): 65P10, 65P40, 65M70, 35L05.

Keywords: Nonlinear wave equation, trigonometric integrators, long-time behaviour, modulated Fourier expansion.

1 Introduction

We consider the nonlinear wave equation

$$\partial_{tt}u - \partial_{xx}u + \rho u = u^2, \quad u = u(x, t) \in \mathbb{R}, \quad (1)$$

on a one-dimensional torus, $x \in \mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$, with a positive Klein–Gordon parameter ρ . We assume that the initial value consists of a single Fourier mode and is small. The smallness of the initial value corresponds to a weakly nonlinear setting. In [15], it has been investigated how the energy, which is initially located in this single Fourier mode, is distributed among the other modes in the course of time. More precisely, the formation of energy strata has been shown that are persist on long time intervals, see Section 2 for a precise description of this result.

In the present paper, we discretize the nonlinear wave equation and answer the question whether this long-time property of the exact solution is inherited by the numerical method. As a numerical method, we consider a spectral collocation in space combined with a symplectic trigonometric integrator in time. We show that this numerical method in fact reproduces the energy strata of the exact solution even

¹Institut für Mathematik, TU Berlin, Straße des 17. Juni 136, D-10623 Berlin, Germany. Present address: Institut für Mathematik, FU Berlin, Arnimallee 9, D-14195 Berlin, Germany (gauckler@math.fu-berlin.de).

²Institut für Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie (KIT), Englerstr. 2, D-76131 Karlsruhe, Germany (daniel.weiss@kit.edu).

on a long time interval, provided that a certain numerical non-resonance condition is fulfilled, see Section 2 for a formulation of this main result.

The considered numerical method is already known to behave well on long time intervals with respect to preservation of regularity and near-conservation of actions [6, 12], as well as near-conservation of energy and momentum [2, 3, 6, 11]. Our result adds to this list an even more sophisticated long-time property of the exact solution that is reproduced in a qualitatively correct way. In comparison to previous results, the considered situation requires less control of interactions in the nonlinearity, which allows us to exclude numerical resonances under weaker restrictions on the time step-size than needed previously.

The proof of our main result is given in Sections 3–5. As for the exact solution, it is based on a modulated Fourier expansion in time, with a multitude of additional difficulties due to the discrete setting that will be described in detail.

Related questions have been studied for symplectic and non-symplectic methods applied to the nonlinear Schrödinger equation. In this case, an initial value consisting of a single Fourier mode yields a solution that continues to consist of a single Fourier mode for all times (plane wave solution). In particular and in contrast to the nonlinear wave equation considered in the present paper, there is no formation of energy strata in the other modes. The stability in numerical discretizations of these plane wave solutions under small perturbations of the initial value is an old question [22] and has been studied on short time intervals [4, 7, 19, 20, 22] and on long time intervals [10].

2 Statement of the main result

2.1 Metastable energy strata revisited

Writing the solution $u = u(x, t)$ of the nonlinear wave equation (1) as a Fourier series $\sum_{j \in \mathbb{Z}} u_j(t) e^{ijx}$ with Fourier coefficients u_j , the *mode energies* of the nonlinear wave equation (1) are given by

$$E_j(t) = \frac{1}{2} |\omega_j u_j(t)|^2 + \frac{1}{2} |\dot{u}_j(t)|^2, \quad j \in \mathbb{Z}, \quad (2)$$

where ω_j are the frequencies

$$\omega_j = \sqrt{j^2 + \rho}, \quad j \in \mathbb{Z}. \quad (3)$$

Note that $E_j = E_{-j}$ for real-valued initial values (and hence real-valued solutions).

We are interested in the evolution of these mode energies. Assuming that the initial values are small and concentrated in the first mode,

$$E_1(0) \leq \epsilon \ll 1, \quad E_j(0) = 0 \quad \text{for} \quad |j| \neq 1,$$

an inspection of interaction of Fourier modes in the nonlinearity $g(u) = u^2$ suggests that the energy in the first mode is distributed among the other modes in a geometrically decaying way:

$$E_0(t) = \mathcal{O}(\epsilon^2), \quad E_j(t) = \mathcal{O}(\epsilon^{|j|}) \quad \text{for} \quad j \neq 0,$$

at least for small times t .

The main result of [15] (Theorem 1 for $\theta = 1$) states that these energy strata are in fact stable on long time intervals in the following sense: for fixed but arbitrary

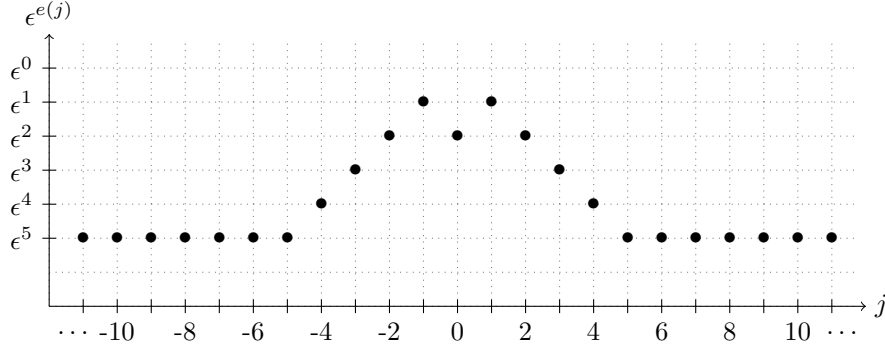


Figure 1: Illustration of the bound $\epsilon^{e(j)}$ for E_j in the case $K = 5$.

$K \geq 2$ and $s > \frac{1}{2}$, we have on the long time interval $0 \leq t \leq c\epsilon^{-K/2}$

$$\begin{aligned} E_0(t) &\leq C\epsilon^2, \\ E_l(t) &\leq C\epsilon^l, \quad 0 < l < K, \\ \sum_{l=K}^{\infty} \sigma_l \epsilon^{-K} E_l(t) &\leq C \end{aligned}$$

with the weights

$$\sigma_j = \max(|j|, 1)^{2s}. \quad (4)$$

This can be written in compact and equivalent form as

$$\sum_{l=0}^{\infty} \sigma_l \epsilon^{-e(l)} E_l(t) \leq C \quad \text{for} \quad 0 \leq t \leq c\epsilon^{-K/2}$$

with the energy profile

$$e(j) = \begin{cases} 2, & j = 0, \\ |j|, & 0 < |j| < K, \\ K, & |j| \geq K, \end{cases} \quad (5)$$

see also Figure 1 for an illustration. In addition to these bounds on all mode energies, the first mode energy is nearly conserved on this long time interval, $|E_1(t) - E_1(0)| \leq C\epsilon^2$. All constants depend on K and s but not on ϵ , and they deteriorate as K or s grow. This result is valid for all but finitely many values of the parameter $\rho \in [0, \rho_0]$ in the wave equation (1), which excludes some resonant situations. The question that we want to answer in the present paper is whether this stable behaviour on long time intervals is inherited by a typical structure-preserving numerical discretization of the nonlinear wave equation.

2.2 Trigonometric integrators

For the numerical discretization of the nonlinear wave equation (1) we consider trigonometric integrators in time applied to a spectral collocation in space, see, for example, [1–3, 6, 9, 11, 12, 14].

In these methods, the solution $u = u(x, t)$ of (1) is approximated, at discrete times $t_n = n\tau$ with the time step-size τ , by a trigonometric polynomial of degree M :

$$u(x, t_n) \approx \sum_{j=-M}^{M-1} u_j^n e^{ijx}, \quad n = 0, 1, \dots$$

The Fourier coefficients $\mathbf{u}^n = (u_{-M}^n, \dots, u_{M-1}^n)^T$ of this trigonometric polynomials are computed with the trigonometric integrator

$$\mathbf{u}^{n+1} - 2 \cos(\tau \mathbf{\Omega}) \mathbf{u}^n + \mathbf{u}^{n-1} = \tau^2 \mathbf{\Psi}((\mathbf{\Phi} \mathbf{u}^n) * (\mathbf{\Phi} \mathbf{u}^n)), \quad (6a)$$

where $\mathbf{\Omega}$ denotes the diagonal matrix containing the frequencies ω_j , $j = -M, \dots, M-1$, of (3). In addition, $*$ denotes the discrete convolution defined by

$$(\mathbf{u} * \mathbf{v})_j = \sum_{j_1 + j_2 \equiv j \pmod{2M}} u_{j_1} v_{j_2}, \quad j = -M, \dots, M-1,$$

which can be computed efficiently using the fast Fourier transform. The method is characterized by the diagonal filter matrices $\mathbf{\Psi} = \psi(\tau \mathbf{\Omega})$ and $\mathbf{\Phi} = \phi(\tau \mathbf{\Omega})$, which are computed from filter functions ψ and ϕ . These filter functions are assumed to be real-valued, bounded and even with $\psi(0) = \phi(0) = 1$. The starting approximation is computed by

$$\mathbf{u}^1 = \cos(\tau \mathbf{\Omega}) \mathbf{u}^0 + \tau \operatorname{sinc}(\tau \mathbf{\Omega}) \dot{\mathbf{u}}^0 + \frac{1}{2} \tau^2 \mathbf{\Psi}((\mathbf{\Phi} \mathbf{u}^0) * (\mathbf{\Phi} \mathbf{u}^0)), \quad (6b)$$

and a velocity approximation by

$$2\tau \operatorname{sinc}(\tau \mathbf{\Omega}) \dot{\mathbf{u}}^n = \mathbf{u}^{n+1} - \mathbf{u}^{n-1}. \quad (6c)$$

An error analysis of these methods has been given in [14]. We assume here that the considered trigonometric integrator is symplectic,

$$\psi(\xi) = \operatorname{sinc}(\xi) \phi(\xi), \quad (7)$$

see [18, Chap. XIII, Eq. (2.10)]. Examples for such filter functions are $\phi = 1$ and $\psi = \operatorname{sinc}$, which is the impulse method [16, 21] or method of Deuffhard [8], as well as $\phi = \operatorname{sinc}$ and $\psi = \operatorname{sinc}^2$, which is the mollified impulse method of García-Archilla, Sanz-Serna & Skeel [13]. Certain choices of filter functions leading to non-symplectic methods could also be handled using the transformation indicated in [5, Remark 3.2], but we do not pursue this here.

We are interested in the mode energies (2) along the numerical solution (6). We denote them in the following by

$$E_j^n = \frac{1}{2} |\omega_j u_j^n|^2 + \frac{1}{2} |\dot{u}_j^n|^2, \quad j = -M, \dots, M-1. \quad (8)$$

We note that $E_j^n = E_{-j}^n$ for real-valued initial values (and hence numerical solutions that take real values in the collocation points $x_k = \pi k/M$, $k = -M, \dots, M-1$), and we set $E_M^n = E_{-M}^n$. As for the exact solution in Section 2.1, we consider initial values with

$$E_1^0 \leq \epsilon, \quad E_j^0 = 0 \quad \text{for} \quad |j| \neq 1. \quad (9)$$

2.3 Metastable energy strata in trigonometric integrators

We now state our main result which says, roughly speaking, that trigonometric integrators (6) integrate the metastable energy strata in nonlinear wave equations (1) qualitatively correctly.

For this result, we need a non-resonance condition on the time step-size τ and on the frequencies ω_j , $j = -M, \dots, M-1$, of (3). In the statement of this condition, we consider indices $j \in \{-M, \dots, M-1\}$ and vectors $\mathbf{k} = (k_0, \dots, k_M)^T$ of integers k_l , and we write

$$\mathbf{k} \cdot \boldsymbol{\omega} = \sum_{l=0}^M k_l \omega_l$$

with the vector $\boldsymbol{\omega} = (\omega_0, \dots, \omega_M)^T$ of frequencies. We fix $3 \leq K \leq M$, and we denote, corresponding to [15], by \mathcal{K} the set

$$\begin{aligned} \mathcal{K} = & \{ (j, \mathbf{k}) : \max(|j|, \mu(\mathbf{k})) < 2K \text{ and } k_l = 0 \text{ for all } l \geq K \} \\ & \cup \{ (j, \pm \langle (j-r) \bmod 2M \rangle + \mathbf{k}) : |(j-r) \bmod 2M| \geq K, |r| < K, \mu(\mathbf{k}) < K \}, \end{aligned} \quad (10)$$

where $\langle j \rangle = (0, \dots, 0, 1, 0, \dots, 0)^T$ is the $|j|$ th unit vector and

$$\mu(\mathbf{k}) = \sum_{l=0}^M |k_l| e(l) = 2|k_0| + \sum_{l=1}^{K-1} |k_l| l + K \sum_{l=K}^M |k_l|. \quad (11)$$

In comparison to the set \mathcal{K} for the exact solution (see [15, Equation (11)]), we have to consider indices modulo $2M$ due to the discretization in space. In addition, we correct here a typo in [15, Equation (11)], where $|j| \geq K$ should be replaced by $|j-r| \geq K$.

Non-resonance condition. For given $0 \leq \nu < \frac{1}{2}$, we assume that there exists a constant $0 < \gamma \leq 1$ such that

$$\left| \sin\left(\frac{1}{2}\tau(\omega_j \pm \mathbf{k} \cdot \boldsymbol{\omega})\right) \right| \geq \gamma \tau \epsilon^{\nu/2} \quad \text{for all} \quad (j, \mathbf{k}) \in \mathcal{K}, \mathbf{k} \neq \mp \langle j \rangle \quad (12a)$$

and

$$\left| \sin\left(\frac{1}{2}\tau(\omega_j \pm \mathbf{k} \cdot \boldsymbol{\omega})\right) \right| \geq \gamma \tau \quad \text{for all} \quad (j, \mathbf{k}) \in \mathcal{K}, \mathbf{k} \neq \langle j \rangle, \mathbf{k} \neq -\langle j \rangle, |j| \leq K. \quad (12b)$$

We emphasize that only $\mathbf{k} = -\langle j \rangle$ is excluded in the estimate for $\sin(\frac{1}{2}\tau(\omega_j + \mathbf{k} \cdot \boldsymbol{\omega}))$ in the first condition (12a), whereas $\mathbf{k} = \langle j \rangle$ and $\mathbf{k} = -\langle j \rangle$ are both excluded in the estimate for this sine in the second condition (12b).

Under this non-resonance condition, we prove in this paper the following discrete analogon of the analytical result [15, Theorem 1] described in Section 2.1.

Theorem 1. *Fix an integer $3 \leq K \leq M$ and real numbers $0 \leq \nu < \frac{1}{2}$ and $s > \frac{1}{2}$, and assume that the non-resonance condition (12) holds for this ν . Then there exist $\epsilon_0 > 0$ and positive constants c and C such that the mode energies (8) along trigonometric integrators (6) for nonlinear wave equations (1) with initial data (9) with $0 < \epsilon \leq \epsilon_0$ satisfy, over long times*

$$0 \leq t_n = n\tau \leq c\epsilon^{-K(1-2\nu)/2},$$

the bounds

$$\sum_{l=0}^M \sigma_l \epsilon^{-e(l)} E_l^n \leq C$$

and $\epsilon^{-e(1)} |E_1^n - E_1^0| \leq C\epsilon^{1-2\nu}$. The constants c and C depend on K , ν and s , but not on ϵ and the discretization parameters τ and M .

Remark 2. The result of Theorem 1 gets stronger for larger K (with worse constants, however), but also the assumption gets stronger, because the set \mathcal{K} (see (10)) in the non-resonance condition (12) grows. We can thus also cover the (not so interesting) border case $K = 2$ as considered in [15] provided that the non-resonance assumption holds for $K = 3$. Without requiring the non-resonance condition with $K = 3$, our techniques of proof can be used to show that Theorem 1 holds for $K = 2$, if we replace $\epsilon^{-e(1)}|E_1^n - E_1^0| \leq C\epsilon^{1-2\nu}$ by $\epsilon^{-e(1)}|E_1^n - E_1^0| \leq C\epsilon^{(1-2\nu)/2}$. The same remark applies to [15].

The proof of Theorem 1 is given in Sections 3–5 below. The structure of the proof is similar to the structure of the corresponding proof for the exact solution given in [15], with a multitude of additional difficulties due to the discrete setting.

The main difference of this result in comparison with the corresponding result [15, Theorem 1] for the exact solution is the required non-resonance condition. The non-resonance condition (12) excludes two types of resonances. First, it requires that $\omega_j \pm \mathbf{k} \cdot \boldsymbol{\omega}$ is bounded away from zero for $(j, \mathbf{k}) \in \mathcal{K}$ with $\mathbf{k} \neq \mp \langle j \rangle$. This is the non-resonance condition for the exact solution, which can be shown to hold for all except finitely many values of ρ in a fixed interval, see [15, Section 3.2]. In addition, the non-resonance condition (12) requires also that products $\tau(\omega_j \pm \mathbf{k} \cdot \boldsymbol{\omega})$ are bounded away from *nonzero* integer multiples of 2π . If this latter condition is violated, we observe numerical resonances as illustrated in the following section.

Numerical resonances can be typically excluded under some CFL-type step-size restriction on the time step-size. A feature of the considered situation is that they can be excluded under less restrictive assumptions than in previous studies on the long-time behaviour of trigonometric integrators for nonlinear wave equations [6, 12]. In particular, there are no numerical resonances on the time interval $0 \leq t \leq \epsilon^{-K/2}$ under the CFL-type step-size restriction

$$\tau(M + K)\sqrt{1 + \rho} \leq c < \pi, \quad (13)$$

for some constant c (we may take $\nu = 0$ here). This follows from the structure of the set \mathcal{K} of (10) which enters the non-resonance condition: this set allows only a single large frequency $\omega_l \leq M\sqrt{1 + \rho}$ in the linear combination $\mathbf{k} \cdot \boldsymbol{\omega}$, and it allows to bound the remaining part of $\mathbf{k} \cdot \boldsymbol{\omega}$ by $2K\sqrt{1 + \rho}$ (using the definition of μ and the bounds on μ in \mathcal{K}). Under the above CFL condition (13) and using $\omega_j \leq M\sqrt{1 + \rho}$, the arguments $\frac{1}{2}\tau(\omega_j \pm \mathbf{k} \cdot \boldsymbol{\omega})$ of the sines in the non-resonance condition (12) are thus bounded by $c < \pi$, which excludes numerical resonances (but not analytical resonances).

In the situation of [6, 12], the stronger CFL-type step-size restriction of the form $\tau MK \leq c < \pi$ has to be used to exclude numerical resonances.

For nonzero ν , the numerical non-resonance condition (12) resembles the one used in [6], see Equations (23) and (24) therein. We note that the reduction (24) there is of no relevance in our context because of the structure of the set \mathcal{K} . We also note that the number of indices (j, \mathbf{k}) that have to satisfy the stronger condition (12b), which does not appear in [6], is bounded independently of the spatial discretization parameter M .

There are several possible extensions of Theorem 1. First, we could extend the result, as in [15], to an energy profile e with $e(l) = K + (|l| - K)(1 - \theta)$ instead of $e(l) = K$ for $|l| \geq K$, where $0 < \theta \leq 1$. This shows that also the mode-energies E_l for $l > K$ decay geometrically, but only with a smaller power of ϵ close to 1. As in [15], the time-scale is then restricted to $0 \leq t_n \leq \epsilon^{-\theta K(1-2\nu)/2}$.

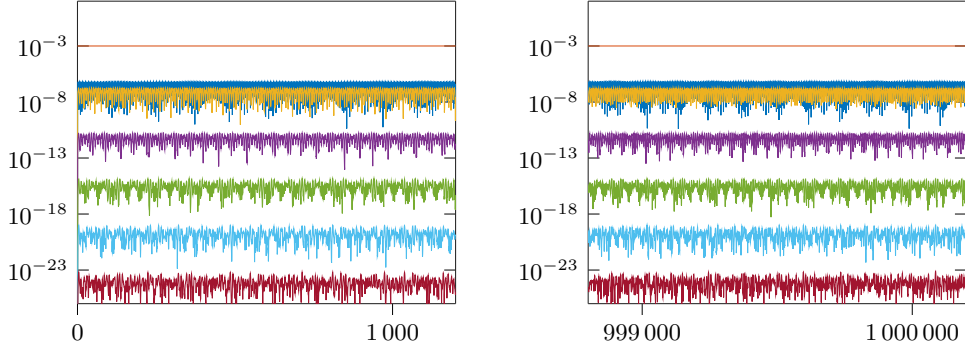


Figure 2: Mode energies E_l^n vs. time t_n for the numerical solution with time step-size $\tau = 0.05$.

Second, the result holds for general nonlinearities $g(u)$ instead of u^2 if g is real-analytic near 0 and $g(0) = g'(0) = 0$. In addition, stronger estimates hold if further derivatives of g vanish at 0.

Finally, the result can be extended to more general initial energy profiles, for example to the situation where a whole band E_l^0 , $|l| \leq B$, of initial mode energies is of order ϵ .

2.4 Numerical experiment

We consider the nonlinear wave equation (1) with $\rho = \sqrt{3}$ and with initial value satisfying (9) for $\epsilon = 10^{-3}$. We apply the trigonometric integrator (6) with filter functions $\phi = 1$ and $\psi = \text{sinc}$ (the impulse method or method of Deuffhard) to the equation. We take $M = 2^5$ for the spectral collocation in space, and we use three different time step-sizes for the discretization in time¹.

The first time step-size is $\tau = 0.05$. For this time step-size, the trigonometric integrator integrates the geometrically decaying energy strata qualitatively correctly, even on long time intervals, see Figure 2 (every tenth time step is plotted there). The time step-size fulfills the CFL-type condition (13) with $K = 6$, under which numerical resonances in the non-resonance condition (12) of Theorem 1 can be excluded.

The second time step-size is $\tau = 2\pi/(\omega_1 + \omega_6 + \omega_7) \approx 0.4212$. It is chosen in such a way that it does not satisfy the non-resonance condition (12). The resulting numerical resonance becomes apparent in Figure 3, where we observe in particular an exchange among the first, sixth and seventh mode. For better visibility, we have plotted $\max_{m=0,\dots,99} E_l^{n+m}$ for $n = 0, 100, 200, \dots$ instead of E_l^n for $n = 0, 1, 2, \dots$ in Figure 3. In the same figure, a similar behaviour can be observed for the resonant time step-size $\tau = 2\pi/(-\omega_1 + \omega_6 + \omega_7)$.

3 Modulated Fourier expansions: Proof of Theorem 1

We assume that the non-resonance conditions (12) holds for some $0 \leq \nu < \frac{1}{2}$.

¹The code is available at http://www.waves.kit.edu/downloads/CRC1173_Preprint_2016-13_supplement.zip

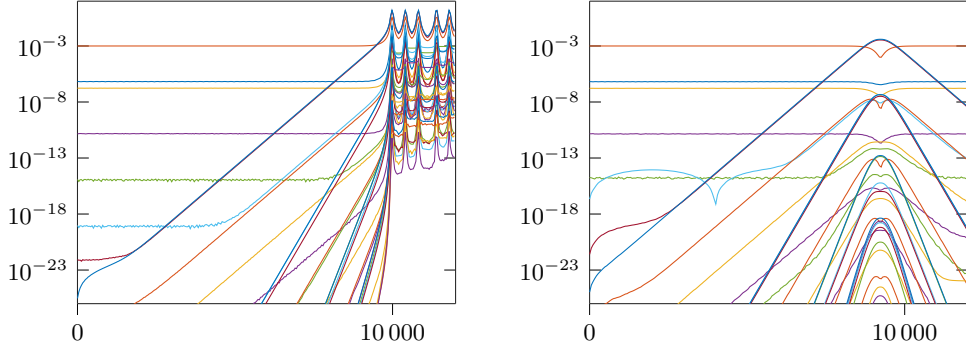


Figure 3: Mode energies E_l^n vs. time t_n for the numerical solutions with time step-size $\tau = 2\pi/(\omega_1 + \omega_6 + \omega_7)$ (left) and time step-size $\tau = 2\pi/(-\omega_1 + \omega_6 + \omega_7)$ (right).

3.1 Approximation ansatz

We will approximate the numerical solution by a *modulated Fourier expansion*,

$$u_j^n \approx \tilde{u}_j^n = \sum_{\mathbf{k} \in \mathcal{K}_j} z_j^{\mathbf{k}}(\epsilon^{\nu/2} t_n) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n}, \quad j = -M, \dots, M-1, \quad (14)$$

where $\mathcal{K}_j = \{\mathbf{k} : (j, \mathbf{k}) \in \mathcal{K}\}$ with the set \mathcal{K} of (10). We require that the *modulation functions* $z_j^{\mathbf{k}}$ are polynomials. In contrast to the modulated Fourier expansion of [15, Equation (7)] for the exact solution, the modulation functions considered here vary on a slow time-scale $\epsilon^{\nu/2} t$, where ν is the parameter from the non-resonance condition (12). With this slow time scale, derivatives of $z_j^{\mathbf{k}}(\epsilon^{\nu/2} t)$ with respect to t carry additional factors $\epsilon^{\nu/2}$, which will be used to compensate for this factor in the weak non-resonance condition (12a).

Having in mind that u_j^n , $j = -M, \dots, M-1$, are the Fourier coefficients of a trigonometric polynomial of degree M , we write in the following $u_M^n = u_{-M}^n$ and $z_M^{\mathbf{k}} = z_{-M}^{\mathbf{k}}$.

Inserting the ansatz (14) into the trigonometric integrator (6a) and comparing the coefficients of $e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t}$, yields a set of equations for the modulation functions $\mathbf{z} = (z_j^{\mathbf{k}})_{(j, \mathbf{k}) \in \mathcal{K}}$, for which solutions are to be constructed up to a small defect $d_j^{\mathbf{k}}$:

$$\begin{aligned} e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) \tau} z_j^{\mathbf{k}}(\epsilon^{\nu/2}(t + \tau)) - 2 \cos(\tau \omega_j) z_j^{\mathbf{k}}(\epsilon^{\nu/2} t) + e^{-i(\mathbf{k} \cdot \boldsymbol{\omega}) \tau} z_j^{\mathbf{k}}(\epsilon^{\nu/2}(t - \tau)) \\ = -\tau^2 \psi(\tau \omega_j) \nabla_{-j}^{-\mathbf{k}} \mathcal{U}(\boldsymbol{\Phi} \mathbf{z}(\epsilon^{\nu/2} t)) + \tau^2 \psi(\tau \omega_j) d_j^{\mathbf{k}}(\epsilon^{\nu/2} t) \end{aligned} \quad (15)$$

with the extended potential

$$\mathcal{U}(\mathbf{z}) = -\frac{1}{3} \sum_{j_1 + j_2 + j_3 \equiv 0} \sum_{\mathbf{k}^1 + \mathbf{k}^2 + \mathbf{k}^3 = \mathbf{0}} z_{j_1}^{\mathbf{k}^1} z_{j_2}^{\mathbf{k}^2} z_{j_3}^{\mathbf{k}^3}, \quad (16)$$

with $\boldsymbol{\Phi} \mathbf{z} = (\phi(\tau \omega_j) z_j^{\mathbf{k}})_{(j, \mathbf{k}) \in \mathcal{K}}$ and with $\nabla_{-j}^{-\mathbf{k}}$ denoting the partial derivative with respect to $z_{-j}^{\mathbf{k}}$. In (16) and in the following we denote by \equiv the congruence modulo $2M$.

With the modulated Fourier expansion (14) at hand, the formula (6c) for the velocity approximation leads to an approximation of \dot{u}_j^n by a modulated Fourier

expansion:

$$\begin{aligned} \dot{u}_j^n \approx \tilde{u}_j^n = (2\tau \operatorname{sinc}(\tau\omega_j))^{-1} \sum_{\mathbf{k} \in \mathcal{K}_j} & \left(z_j^{\mathbf{k}}(\epsilon^{\nu/2} t_{n+1}) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})\tau} \right. \\ & \left. - z_j^{\mathbf{k}}(\epsilon^{\nu/2} t_{n-1}) e^{-i(\mathbf{k} \cdot \boldsymbol{\omega})\tau} \right) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})t_n}. \end{aligned} \quad (17)$$

Finally, we get conditions from the fact that the ansatz (14) should satisfy the initial condition:

$$u_j^0 = \sum_{\mathbf{k} \in \mathcal{K}_j} z_j^{\mathbf{k}}(0), \quad (18a)$$

$$\dot{u}_j^0 = (2\tau \operatorname{sinc}(\tau\omega_j))^{-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \left(z_j^{\mathbf{k}}(\epsilon^{\nu/2}\tau) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})\tau} - z_j^{\mathbf{k}}(-\epsilon^{\nu/2}\tau) e^{-i(\mathbf{k} \cdot \boldsymbol{\omega})\tau} \right), \quad (18b)$$

where (17) was used for the derivation of the second equation.

3.2 Norms

For vectors $\mathbf{v} = (v_{-M}, \dots, v_{M-1})$ of Fourier coefficients of trigonometric polynomials of degree M , we consider the norm

$$\|\mathbf{v}\| = \left(\sum_{j=-M}^{M-1} \sigma_j \epsilon^{-2e(j)\nu} |v_j|^2 \right)^{1/2},$$

where $e(j)$ is the energy profile of (5) and σ_j are the weights of (4). For $\nu = 0$, this is the Sobolev H^s -norm of the corresponding trigonometric polynomial $\sum_{j=-M}^{M-1} v_j e^{ijx}$. The rescaling $\epsilon^{-2e(j)\nu}$ originates from the non-resonance condition (12a) that introduces $\epsilon^{\nu/2}$.

We will make frequent use of the fact that this norm behaves well with respect to the discrete convolution,

$$\|\mathbf{u} * \mathbf{v}\| \leq C \|\mathbf{u}\| \|\mathbf{v}\|. \quad (19)$$

This follows from the corresponding property of the Sobolev H^s -norm, see, for example, [17, Lemma 4.2], and the fact that $e(j) \leq e(j_1) + e(j_2)$ for $j \equiv j_1 + j_2$. Similarly, we also have

$$\|\Omega(\mathbf{u} * \mathbf{v})\| \leq C \|\Omega\mathbf{u}\| \|\Omega\mathbf{v}\|. \quad (20)$$

In particular, we have for $\nu = 0$ in (19)

$$\sum_{j=-M}^{M-1} \sigma_j \left| \sum_{j_1+j_2 \equiv j} u_{j_1} v_{j_2} \right|^2 \leq C^2 \left(\sum_{j_1=-M}^{M-1} \sigma_{j_1} |u_{j_1}|^2 \right) \left(\sum_{j_2=-M}^{M-1} \sigma_{j_2} |v_{j_2}|^2 \right). \quad (21)$$

3.3 The modulated Fourier expansion on a short time interval

Instead of initial mode energies (9), we consider the more general choice

$$\sum_{l=0}^M \sigma_l \epsilon^{-e(l)} E_l^0 \leq C_0 \quad (22)$$

of initial mode energies, which is the expected situation at later times (see Theorem 1). We then have the following discrete counterpart of [15, Theorem 3], whose proof will be given in Section 4.

Theorem 3. Under the assumptions of Theorem 1, but with (22) instead of (9), the numerical solution $\mathbf{u}^n = (u_{-M}^n, \dots, u_{M-1}^n)^\top$ admits an expansion

$$u_j^n = \sum_{\mathbf{k} \in \mathcal{K}_j} z_j^{\mathbf{k}}(\epsilon^{\nu/2} t_n) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n} + r_j^n \quad \text{for} \quad 0 \leq t_n = n\tau \leq 1,$$

where $\mathcal{K}_j = \{ \mathbf{k} : (j, \mathbf{k}) \in \mathcal{K} \}$ with the set \mathcal{K} of (10), where the coefficient functions $z_j^{\mathbf{k}}$ are polynomials satisfying $z_{-j}^{-\mathbf{k}} = \overline{z_j^{\mathbf{k}}}$, and where the remainder term $\mathbf{r}^n = (r_{-M}^n, \dots, r_{M-1}^n)^\top$ and the corresponding remainder term $\dot{\mathbf{r}}^n = (\dot{r}_{-M}^n, \dots, \dot{r}_{M-1}^n)^\top$ in the velocity approximation (17) are bounded by

$$\|\Omega \mathbf{r}^n\| + \|\dot{\mathbf{r}}^n\| \leq C \epsilon^{K(1-2\nu)} \quad \text{for} \quad 0 \leq t_n = n\tau \leq 1$$

and the defect $d_j(t) = \sum_{\mathbf{k} \in \mathcal{K}_j} |d_j^{\mathbf{k}}(\epsilon^{\nu/2} t)|$ in (15) is bounded by

$$\|\mathbf{d}(t)\| \leq C \epsilon^{K(1-2\nu)} \quad \text{for} \quad 0 \leq t \leq 1.$$

The constant C is independent of ϵ , τ and M , but depends on K , on ν and γ of (12), on ρ of (1), on s of (4) and on C_0 of (22).

3.4 Almost-invariant energies

We now derive almost-invariant energies of the modulation system (15) which enable us to consider long time intervals. The derivation of these almost-invariant energies is similar as in the case of the exact solution, see [15, Section 3.5], but it now leads to discrete almost-invariant energies that involve additionally the time step-size τ and are related to those of [6].

Let

$$\mathcal{E}_l(t) = -\frac{i}{2} \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \frac{k_l \omega_l}{\tau \operatorname{sinc}(\tau \omega_j)} z_{-j}^{-\mathbf{k}}(\epsilon^{\nu/2} t) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) \tau} z_j^{\mathbf{k}}(\epsilon^{\nu/2} (t + \tau)). \quad (23)$$

Our aim is to prove that

$$\mathcal{E}_l(t) = \mathcal{E}_l(t - \tau) - \frac{i}{2} \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \tau k_l \omega_l \phi(\tau \omega_j) z_{-j}^{-\mathbf{k}}(\epsilon^{\nu/2} t) d_j^{\mathbf{k}}(\epsilon^{\nu/2} t), \quad (24)$$

which shows that, for small defect $d_j^{\mathbf{k}}$, the quantity \mathcal{E}_l is almost invariant. To see this almost-invariance, we use, as in the case of the exact solution, that the extended potential \mathcal{U} of (16) is invariant under the transformation $\mathcal{S}_{\boldsymbol{\lambda}}(\theta) \mathbf{z} := (e^{i(\mathbf{k} \cdot \boldsymbol{\lambda}) \theta} z_j^{\mathbf{k}})_{(j, \mathbf{k}) \in \mathcal{K}}$ for $\theta \in \mathbb{R}$ and real vectors $\boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_M)$. This invariance shows that

$$0 = \frac{d}{d\theta} \Big|_{\theta=0} \mathcal{U}(\mathcal{S}_{\boldsymbol{\lambda}}(\theta) \boldsymbol{\Phi} \mathbf{z}) = - \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} i(\mathbf{k} \cdot \boldsymbol{\lambda}) \phi(\tau \omega_j) z_{-j}^{-\mathbf{k}} \nabla_{-j}^{-\mathbf{k}} \mathcal{U}(\boldsymbol{\Phi} \mathbf{z}).$$

We then multiply this equation with $\tau/2$, we replace $-\nabla_{-j}^{-\mathbf{k}} \mathcal{U}$ with the help of (15), and we use the symplecticity (7) of the method. This yields

$$\begin{aligned} 0 = & \frac{i}{2} \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \frac{\mathbf{k} \cdot \boldsymbol{\lambda}}{\tau \operatorname{sinc}(\tau \omega_j)} z_{-j}^{-\mathbf{k}}(\epsilon^{\nu/2} t) \left(e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) \tau} z_j^{\mathbf{k}}(\epsilon^{\nu/2} (t + \tau)) \right. \\ & \left. - 2 \cos(\tau \omega_j) z_j^{\mathbf{k}}(\epsilon^{\nu/2} t) + e^{-i(\mathbf{k} \cdot \boldsymbol{\omega}) \tau} z_j^{\mathbf{k}}(\epsilon^{\nu/2} (t - \tau)) - \tau^2 \psi(\tau \omega_j) d_j^{\mathbf{k}}(\epsilon^{\nu/2} t) \right). \end{aligned}$$

In this relation, we use that

$$\sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \frac{\mathbf{k} \cdot \boldsymbol{\lambda}}{\text{sinc}(\tau\omega_j)}, z_{-j}^{-\mathbf{k}}(\epsilon^{\nu/2}t) \cos(\tau\omega_j) z_j^{\mathbf{k}}(\epsilon^{\nu/2}t) = 0, \quad (25)$$

which follows from the symmetry in j (recall that $z_M^{\mathbf{k}} = z_{-M}^{\mathbf{k}}$) and the asymmetry in \mathbf{k} . Choosing $\boldsymbol{\lambda} = \langle l \rangle$ with $l = 0, \dots, M$ and using the definition (23) of \mathcal{E}_l , we then end up with (24).

Note that, due to the appearance of the factor k_l in \mathcal{E}_l of (23), only modulation functions $z_j^{\mathbf{k}}$ with $k_l \neq 0$ are actually relevant in \mathcal{E}_l (recall that k_l is the l th entry of the vector \mathbf{k}). This will be used all the time in Section 5 below. In particular, we will use there that only very specific vectors \mathbf{k} with $k_l \neq 0$ for $l \geq K$ appear in the set \mathcal{K} of (10).

Using a Taylor expansion of $z_j^{\mathbf{k}}(\epsilon^{\nu/2}(t + \tau))$ and the property (25), we can derive the following alternative form of \mathcal{E}_l :

$$\begin{aligned} \mathcal{E}_l(t) = & \frac{1}{2} \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \left(k_l \omega_l(\mathbf{k} \cdot \boldsymbol{\omega}) \frac{\text{sinc}(\tau(\mathbf{k} \cdot \boldsymbol{\omega}))}{\text{sinc}(\tau\omega_j)} |z_j^{\mathbf{k}}(\epsilon^{\nu/2}t)|^2 \right. \\ & \left. - i \frac{k_l \omega_l}{\text{sinc}(\tau\omega_j)} z_{-j}^{-\mathbf{k}}(\epsilon^{\nu/2}t) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})\tau} \left(\epsilon^{\nu/2} \dot{z}_j^{\mathbf{k}}(\epsilon^{\nu/2}t) + \frac{1}{2} \tau \epsilon^{\nu} \ddot{z}_j^{\mathbf{k}}(\epsilon^{\nu/2}t) + \dots \right) \right). \end{aligned} \quad (26)$$

Using (25), we see that this quantity coincides, for $\nu = 0$ and in the limits $\tau \rightarrow 0$ and $M \rightarrow \infty$, with the almost-invariant energy of [15, Equation (23)] for the exact solution. We therefore also call $\mathcal{E}_l(t)$ an *almost-invariant energy*. In Sections 4 and 5 below, we prove the following discrete counterparts of [15, Theorems 4–7] for this new almost-invariant energy.

Theorem 4 (Almost-invariant energies controlled by mode energies). *Under the conditions of Theorem 3 we have, for $0 \leq t \leq 1$,*

$$\sum_{l=0}^M \sigma_l \epsilon^{-e(l)} |\mathcal{E}_l(t)| \leq C_0, \quad (27)$$

where C_0 is independent of ϵ , τ and M and depends on the initial data only through the constant C_0 of (22).

Theorem 5 (Variation of almost-invariant energies). *Under the conditions of Theorem 3 we have, for $0 \leq t_n \leq 1$,*

$$\sum_{l=0}^M \sigma_l \epsilon^{-e(l)} |\mathcal{E}_l(t_n) - \mathcal{E}_l(0)| \leq C \epsilon^{K(1-2\nu)/2}$$

and $\epsilon^{-e(1)} |\mathcal{E}_1(t_n) - \mathcal{E}_1(0)| \leq C \epsilon^{(K-1)(1-2\nu)/2} \epsilon^{K(1-2\nu)/2}$, where C is independent of ϵ , τ and M and depends on the initial data only through the constant C_0 of (22).

At time $t = 1$, for which we assume without loss of generality that $t_N = N\tau = 1$ for some N , we consider a new modulated Fourier expansion leading to new almost-invariant energies.

Theorem 6 (Transitions in the almost-invariant energies). *Let the conditions of Theorem 3 be fulfilled. Let $\mathbf{z}(\epsilon^{\nu/2}t) = (z_j^{\mathbf{k}}(\epsilon^{\nu/2}t))_{(j,\mathbf{k}) \in \mathcal{K}}$ for $0 \leq t \leq 1$ be the coefficient*

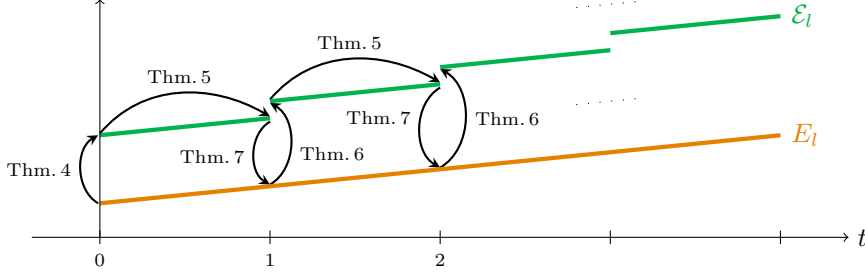


Figure 4: A schematic overview of the proof of Theorem 1: How the results on the almost-invariant energies \mathcal{E}_l are applied to control the mode energies E_l and vice versa.

functions as in Theorem 3 with corresponding almost-invariant energies $\mathcal{E}_l(t)$ for initial data $(\mathbf{u}^0, \dot{\mathbf{u}}^0)$. Let further $\tilde{\mathbf{z}}(\epsilon^{\nu/2}t) = (\tilde{z}_j^{\mathbf{k}}(\epsilon^{\nu/2}t))_{(j,\mathbf{k}) \in \mathcal{K}}$ be the coefficient functions and $\tilde{\mathcal{E}}(t)$ the corresponding almost-invariants of the modulated Fourier expansion for $0 \leq t \leq 1$ corresponding to initial data $(\mathbf{u}^N, \dot{\mathbf{u}}^N)$ with $t_N = N\tau = 1$, constructed as in Theorem 3. If $(\mathbf{u}^N, \dot{\mathbf{u}}^N)$ also satisfies the bound (22), then

$$\sum_{l=0}^M \sigma_l \epsilon^{-e(l)} |\mathcal{E}_l(1) - \tilde{\mathcal{E}}_l(0)| \leq C \epsilon^{K(1-2\nu)/2}$$

and $\epsilon^{-e(1)} |\mathcal{E}_1(1) - \tilde{\mathcal{E}}_1(0)| \leq C \epsilon^{(K-1)(1-2\nu)/2} \epsilon^{K(1-2\nu)/2}$, where C is independent of ϵ , τ and M and depends on the initial data only through the constant C_0 of (22).

Theorem 7 (Mode energies controlled by almost-invariant energies). *Let the conditions of Theorem 3 be fulfilled. If the almost-invariant energies satisfy (27) for $0 \leq t_n = n\tau \leq 1$, then the mode energies are bounded by*

$$\sum_{l=0}^M \sigma_l \epsilon^{-e(l)} E_l^n \leq C$$

and $\epsilon^{-e(1)} |E_1^n - \mathcal{E}_1(t_n)| \leq C \epsilon^{1-2\nu}$, where C depends on C_0 in (27), but is independent of ϵ , τ , M and C_0 of (22) if $\epsilon^{1-2\nu}$ is sufficiently small.

3.5 From short to long time intervals: Proof of Theorem 1

Based on Theorems 3–7, the proof of Theorem 1 is the same as in the case of the exact solution, see [15, Section 3.6]: Theorems 3–5 and 7 yield the statement of Theorem 1 on a short time interval $0 \leq t_n \leq 1$. Theorem 6 can be used to patch many of these short time intervals together, on which the almost-invariant energies \mathcal{E}_l are still well preserved (Theorems 5 and 6) and allow to control the mode energies E_l (Theorem 7). A schematic overview of the proof is given in Figure 4.

4 Construction of a modulated Fourier expansion: Proofs of Theorems 3 and 4

Throughout this section, we work under the assumptions of Theorem 3. In addition, we let

$$\delta = \epsilon^{1/2},$$

such that $u_j^0, \dot{u}_j^0 = \mathcal{O}(\delta^{e(l)})$. Moreover, we write \lesssim for an inequality up to a factor that is independent of $\delta = \epsilon^{1/2}$, τ and M .

4.1 Expansion of the modulation functions

The construction of modulation functions $z_j^{\mathbf{k}}$ of the modulated Fourier expansion (14) is based on an expansion

$$\begin{aligned} z_j^{\mathbf{k}} &= \delta^{m(j, \mathbf{k})} \sum_{m=m(j, \mathbf{k})}^{2K-1} \delta^{(m-m(j, \mathbf{k}))(1-2\nu)} z_{j,m}^{\mathbf{k}} \\ &= \delta^{2m(j, \mathbf{k})\nu} \sum_{m=m(j, \mathbf{k})}^{2K-1} \delta^{m(1-2\nu)} z_{j,m}^{\mathbf{k}} \end{aligned} \quad (28)$$

with polynomials $z_{j,m}^{\mathbf{k}} = z_{j,m}^{\mathbf{k}}(\delta^\nu t)$ and with

$$m(j, \mathbf{k}) = \max\left(e(j), \max_{l: k_l \neq 0} e(l)\right).$$

This can be motivated by the fact that on the one hand we expect $z_j^{\mathbf{k}} = \mathcal{O}(\delta^{m(j, \mathbf{k})})$ from the analysis of the exact solution in [15], but on the other hand we do not expect an expansion in true powers of δ as in [15] because of the non-resonance condition (12) involving δ^ν .

Our goal now is to derive equations for the modulation coefficient functions $z_{j,m}^{\mathbf{k}}$ in the ansatz (28).

Inserting this ansatz into (15), expanding $z_j^{\mathbf{k}}(\delta^\nu(t \pm \tau))$ in a Taylor series and requiring that powers of $\delta^{1-2\nu}$ agree on both sides yields, neglecting the defect d_j^k ,

$$\begin{aligned} 4s_{\langle j \rangle - \mathbf{k}} s_{\langle j \rangle + \mathbf{k}} z_{j,m}^{\mathbf{k}} + 2i\tau\delta^\nu s_{2\mathbf{k}} \dot{z}_{j,m}^{\mathbf{k}} + \tau^2\delta^{2\nu} c_{2\mathbf{k}} \ddot{z}_{j,m}^{\mathbf{k}} + \dots \\ = \tau^2 \psi(\tau\omega_j) \sum_{\mathbf{k}^1 + \mathbf{k}^2 = \mathbf{k}} \sum_{j_1 + j_2 \equiv j} \delta^{2(m(j_1, \mathbf{k}^1) + m(j_2, \mathbf{k}^2) - m(j, \mathbf{k}))\nu} \\ \sum_{m_1 + m_2 = m} \phi(\tau\omega_{j_1}) z_{j_1, m_1}^{\mathbf{k}^1} \phi(\tau\omega_{j_2}) z_{j_2, m_2}^{\mathbf{k}^2}. \end{aligned} \quad (29)$$

Here, we use the notation $s_{\mathbf{k}} = \sin(\frac{\tau}{2}\mathbf{k} \cdot \boldsymbol{\omega})$ and $c_{\mathbf{k}} = \cos(\frac{\tau}{2}\mathbf{k} \cdot \boldsymbol{\omega})$ and the fact that $\cos x - \cos y = 2 \sin(\frac{y-x}{2}) \sin(\frac{y+x}{2})$. All functions in (29) are evaluated at $\delta^\nu t$ and dots on the functions $z_{j,m}^{\mathbf{k}}$ denote derivatives with respect to the slow time scale $\delta^\nu t$. Note that $m(j, \mathbf{k}) \leq m(j_1, \mathbf{k}^1) + m(j_2, \mathbf{k}^2)$ if $\mathbf{k} = \mathbf{k}^1 + \mathbf{k}^2$ and $j \equiv j_1 + j_2 \pmod{2M}$, and hence the power of δ on the right-hand side of (29) is small. We recall that \equiv denotes the congruence modulo $2M$.

Remark 8. For $\nu = 0$ we recover in (29), after division by τ^2 and in the limit $\tau \rightarrow 0$, the system of equations which was used in the case of the exact solution, see [15, Equation (28)]. For $\nu > 0$, the above construction becomes significantly more involved than there. The reason is that the non-resonance condition (12a) only allows us to bound the factor $s_{\langle j \rangle - \mathbf{k}} s_{\langle j \rangle + \mathbf{k}}$ on the left-hand side of (29) from below by $\gamma^2 \tau^2 \delta^{2\nu}$. As we will see in the proof of Lemma 10 below, we can compensate this possibly small factor with an additional $\delta^{2\nu}$ on the right-hand side (together with τ^2), which we gain from the special choice (28) as ansatz for $z_j^{\mathbf{k}}$.

In addition to (29), we get from condition (18) that

$$\sum_{\mathbf{k} \in \mathcal{K}_j} \delta^{2(m(j, \mathbf{k}) - e(j))\nu} z_{j, m}^{\mathbf{k}}(0) = \begin{cases} \delta^{-e(j)} u_j^0, & m = e(j), \\ 0, & \text{else,} \end{cases} \quad (30a)$$

$$\begin{aligned} \sum_{\mathbf{k} \in \mathcal{K}_j} \delta^{2(m(j, \mathbf{k}) - e(j))\nu} & \left(2i s_{2\mathbf{k}} z_{j, m}^{\mathbf{k}}(0) + 2\tau \delta^\nu c_{2\mathbf{k}} \dot{z}_{j, m}^{\mathbf{k}}(0) + i\tau^2 \delta^{2\nu} s_{2\mathbf{k}} \ddot{z}_{j, m}^{\mathbf{k}}(0) + \dots \right) \\ & = \begin{cases} 2\tau \text{sinc}(\tau \omega_j) \delta^{-e(j)} \dot{u}_j^0, & m = e(j), \\ 0, & \text{else,} \end{cases} \end{aligned} \quad (30b)$$

where we use again a Taylor expansion of the modulation functions.

4.2 Construction of modulation functions

We construct polynomial modulation functions of the form (28) by solving (29) and (30) consecutively for $m = 1, 2, \dots, 2K - 1$. For convenience, we set $z_{j, m'}^{\mathbf{k}} = 0$ for $m' < m(j, \mathbf{k})$. Assuming that we have computed all polynomials $z_{j, m'}^{\mathbf{k}}$ for $m' < m$, the construction relies on the observation that only already computed functions $z_{j, m'}^{\mathbf{k}'}$ appear on the right-hand side of (29). This equation is thus, for $z(s) = z_{j, m}^{\mathbf{k}}(s)$, of the form

$$\alpha_0 z(s) - \alpha_1 \dot{z}(s) - \alpha_2 \ddot{z}(s) - \dots - \alpha_L z^{(L)}(s) = p(s) \quad (31)$$

with coefficients $\alpha_0, \dots, \alpha_L$ and a polynomial p .

For $\mathbf{k} \in \mathcal{K}_j$ with $\mathbf{k} \neq \pm \langle j \rangle$, the coefficient α_0 in this equation is nonzero (by the non-resonance condition (12)), and the unique polynomial solution of this equation is given by

$$z(s) = \sum_{k=0}^{\deg(p)} \left(\frac{\alpha_1}{\alpha_0} \frac{d}{ds} + \frac{\alpha_2}{\alpha_0} \frac{d^2}{ds^2} + \dots + \frac{\alpha_L}{\alpha_0} \frac{d^L}{ds^L} \right)^k \frac{1}{\alpha_0} p(s). \quad (32)$$

The modulation coefficient functions constructed in this way are called off-diagonal modulation coefficient functions.

For $\mathbf{k} = \pm \langle j \rangle$, the coefficient α_0 is zero, and the polynomial solutions of (31) are given by

$$z(s) = z(0) + \int_0^s \dot{z}(\sigma) d\sigma \quad (33a)$$

with

$$\dot{z}(s) = \sum_{k=0}^{\deg(p)} \left(\frac{\alpha_2}{\alpha_1} \frac{d}{ds} + \dots + \frac{\alpha_L}{\alpha_1} \frac{d^{L-1}}{ds^{L-1}} \right)^k \frac{(-1)^{k+1}}{\alpha_1} p(s). \quad (33b)$$

In (33a), the initial value $z(0)$ is still a free parameter. We use (30) to fix it. Adding and subtracting the two equations of (30), after multiplying the first equation with $2i\omega_j$ and the second one with $\omega_j/s_{2\langle j \rangle} = 1/(\tau \text{sinc}(\tau \omega_j))$, gives us

$$\begin{aligned} 2i\omega_j z_{j, m}^{\pm \langle j \rangle}(0) &= -i \sum_{\substack{\mathbf{k} \in \mathcal{K}_j \\ \mathbf{k} \neq \langle j \rangle, \mathbf{k} \neq -\langle j \rangle}} \delta^{2(m(j, \mathbf{k}) - e(j))\nu} \left(\omega_j \pm (\mathbf{k} \cdot \boldsymbol{\omega}) \frac{\text{sinc}(\tau(\mathbf{k} \cdot \boldsymbol{\omega}))}{\text{sinc}(\tau \omega_j)} \right) z_{j, m}^{\mathbf{k}}(0) \\ &\mp \sum_{\mathbf{k} \in \mathcal{K}_j} \frac{\delta^{2(m(j, \mathbf{k}) - e(j))\nu}}{\text{sinc}(\tau \omega_j)} \left(\delta^\nu c_{2\mathbf{k}} \dot{z}_{j, m}^{\mathbf{k}}(0) + \frac{i}{2} \tau \delta^{2\nu} s_{2\mathbf{k}} \ddot{z}_{j, m}^{\mathbf{k}}(0) + \dots \right) \\ &+ \delta^{-e(j)} \begin{cases} i\omega_j u_j^0 \pm \dot{u}_j^0, & m = e(j), \\ 0, & \text{else.} \end{cases} \end{aligned} \quad (34)$$

Note again that this equation becomes for $\nu = 0$ and in the limit $\tau \rightarrow 0$ the corresponding equation for the exact solution, see [15, Equation (31)]. The modulation coefficient functions constructed with (33) and (34) are called diagonal modulation coefficient functions.

Example 9. We illustrate the above construction for the first two nontrivial values of m . For simplicity, we restrict to the case $\nu = 0$ and $\phi = 1$ (and hence $\psi = \text{sinc}$ by (7)).

For $m = 1$, we observe that the polynomial p in (31), which is the right-hand side of (29), vanishes for all j and all \mathbf{k} . The above construction (32) thus yields

$$z_{j,1}^{\mathbf{k}} \equiv 0 \quad \text{for} \quad \mathbf{k} \neq \pm\langle j \rangle,$$

and (33b) yields $z_{j,1}^{\pm\langle j \rangle} \equiv 0$. The computation of the initial values for $z_{j,1}^{\pm\langle j \rangle}$ with (34) finally yields

$$z_{j,1}^{\pm\langle j \rangle} \equiv 0 \quad \text{for} \quad |j| \neq 1$$

and

$$z_{1,1}^{\pm\langle 1 \rangle} \equiv \frac{1}{2i\omega_1} \delta^{-1} (i\omega_1 u_1^0 \pm \dot{u}_1^0), \quad z_{-1,1}^{\pm\langle 1 \rangle} \equiv \frac{1}{2i\omega_1} \delta^{-1} (i\omega_1 u_{-1}^0 \pm \dot{u}_{-1}^0)$$

by (33a).

For $m = 2$, we observe that the polynomial p in (31), which is the right-hand side of (29), is non-zero only for $j \in \{\pm 2, 0\}$ and $\mathbf{k} = \{\pm 2\langle 1 \rangle, \mathbf{0}\}$. In particular, we get from (32)

$$z_{j,2}^{\mathbf{k}} \equiv 0 \quad \text{for} \quad \mathbf{k} \neq \pm\langle j \rangle, \quad (j, \mathbf{k}) \notin \{\pm 2, 0\} \times \{\pm 2\langle 1 \rangle, \mathbf{0}\}.$$

For $j = 2$ and $\mathbf{k} = 2\langle 1 \rangle$, the polynomial p is non-zero but constant, and we get from (32) the non-zero off-diagonal modulation coefficient function

$$z_{2,2}^{2\langle 1 \rangle} = \frac{\tau^2 \text{sinc}(\tau\omega_2)}{4s_{\langle 2 \rangle - 2\langle 1 \rangle} s_{\langle 2 \rangle + 2\langle 1 \rangle}} z_{1,1}^{\langle 1 \rangle} \cdot z_{1,1}^{\langle 1 \rangle},$$

and similar expressions for $z_{2,2}^{\mathbf{0}}$, $z_{2,2}^{-2\langle 1 \rangle}$, $z_{-2,2}^{\pm 2\langle 1 \rangle}$, $z_{-2,2}^{\mathbf{0}}$, $z_{0,2}^{\pm 2\langle 1 \rangle}$ and $z_{0,2}^{\mathbf{0}}$. For the diagonal modulation coefficient functions, we get $z_{j,2}^{\pm\langle j \rangle} \equiv 0$ from (33b), and the computation of the initial values with (34) finally yields by (33a)

$$z_{j,2}^{\pm\langle j \rangle} \equiv 0 \quad \text{for} \quad j \notin \{\pm 2, 0\}$$

and

$$z_{2,2}^{\pm\langle 2 \rangle} \equiv \frac{1}{2i\omega_2} \left(-i(\omega_2 \pm 2\omega_1) \frac{\text{sinc}(2\tau\omega_1)}{\text{sinc}(\tau\omega_2)} z_{2,2}^{2\langle 1 \rangle} - i\omega_2 z_{2,2}^{\mathbf{0}} \right. \\ \left. - i(\omega_2 \mp 2\omega_1) \frac{\text{sinc}(-2\tau\omega_1)}{\text{sinc}(\tau\omega_2)} z_{2,2}^{-2\langle 1 \rangle} + \delta^{-2} (i\omega_2 u_2^0 \pm \dot{u}_2^0) \right)$$

and similar expressions for $z_{-2,2}^{\pm\langle -2 \rangle}$ and $z_{0,2}^{\pm\langle 0 \rangle}$.

By means of the expansions (14) and (28) (the latter truncated at $m = 2$ instead of $m = 2K - 1$), we get from the computed modulation coefficient functions $z_{j,m}^{\mathbf{k}}$ with $m = 1, 2$ approximations (recall that $\nu = 0$ here)

$$\widehat{u}_j^n = \sum_{\mathbf{k} \in \mathcal{K}_j} \sum_{m=m(j,\mathbf{k})}^2 \delta^m z_{j,m}^{\mathbf{k}}(t_n) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n}$$

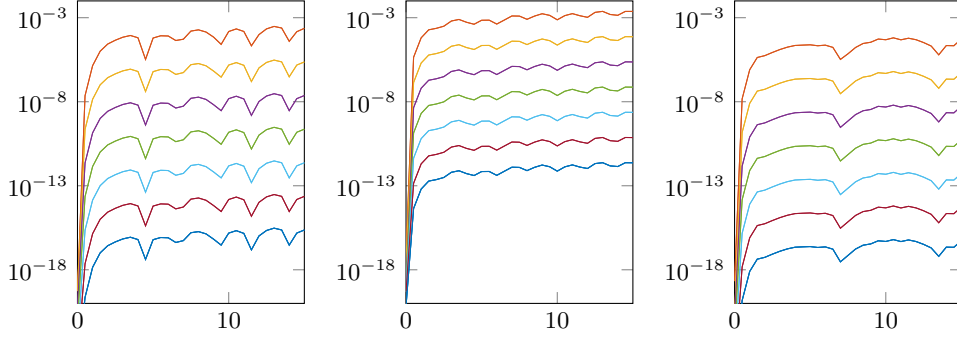


Figure 5: Approximation error $|u_j^n - \hat{u}_j^n|$ vs. time t_n for $j = 0$ (left), $j = 1$ (middle) and $j = 2$ (right). Different lines correspond to different values of $\epsilon = 10^{-2}, 10^{-3}, \dots, 10^{-8}$ in (9).

to the numerical solution u_j^n . For $\tau = 0.05$ and $\rho = \sqrt{3}$, the quality of this approximation is illustrated in Figure 5 for initial values satisfying (9) with different values of ϵ . We observe that the approximation is exact for $n = 0$ (by construction) and that the approximation is of order δ^3 for $j = 1$ on a time interval of order one. We also observe an improved approximation order of δ^4 for $j = 0$ and $j = 2$. This can be explained with the construction for $m = 3$, which yields $z_{2,3}^{\mathbf{k}} = z_{0,3}^{\mathbf{k}} \equiv 0$ for all \mathbf{k} .

Going beyond this example, we verify as in [15, Section 4.1] the following properties of the constructed functions $z_{j,m}^{\mathbf{k}}$. Using $e(j) \leq e(j_1) + e(j_2)$ for $j \equiv j_1 + j_2$, $\mu(\mathbf{k}) \leq \mu(\mathbf{k}^1) + \mu(\mathbf{k}^2)$ for $\mathbf{k} = \mathbf{k}^1 + \mathbf{k}^2$ and $e(j) = \mu(\mathbf{k}) < \mu(\mathbf{k}^1) + \mu(\mathbf{k}^2)$ for $\mathbf{k} = \pm \langle j \rangle = \mathbf{k}^1 + \mathbf{k}^2$ with μ defined in (11), we see that they are polynomials of degree

$$\deg(z_{j,m}^{\mathbf{k}}) \leq m - \max(e(j), \mu(\mathbf{k})), \quad (35)$$

where a negative degree corresponds to the zero polynomial. Moreover, the polynomial $z_{j,m}^{\mathbf{k}}$ can be different from the zero polynomial only in the two cases

$$\text{case 1: } |j| \leq m, \quad \mu(\mathbf{k}) \leq m, \quad k_l = 0 \text{ for } l \geq \min(m+1, K), \quad (36)$$

$$\text{case 2: } m \geq K, \quad |(j-r) \bmod 2M| \geq K, \quad |r| \leq m-K, \\ \mathbf{k} = \pm \langle (j-r) \bmod 2M \rangle + \bar{\mathbf{k}}, \quad \mu(\bar{\mathbf{k}}) \leq m-K. \quad (37)$$

This explains how the set \mathcal{K} of (10) is built up. The two cases follow from the decomposition

$$2\phi(\tau\omega_{j_1})\phi(\tau\omega_{j_2}) \sum_{l=1}^{m-K} z_{j_1,m-l}^{\mathbf{k}^1} z_{j_2,l}^{\mathbf{k}^2} + \phi(\tau\omega_{j_1})\phi(\tau\omega_{j_2}) \sum_{\substack{m_1+m_2=m \\ m_1 < K, m_2 < K}} z_{j_1,m_1}^{\mathbf{k}^1} z_{j_2,m_2}^{\mathbf{k}^2}$$

of the last line in (29), where the first term is only present for $m > K$: the functions $z_{j_i,m_i}^{\mathbf{k}^i}$ in the second sum and the function $z_{j_2,l}^{\mathbf{k}^2}$ in the first sum belong inductively to the first case above, whereas the function $z_{j_1,m-l}^{\mathbf{k}^1}$ in the first sum belongs either to the first or to the second case, leading to a function $z_{j,m}^{\mathbf{k}}$ belonging to the first or second case, respectively. In addition, we have $\overline{z_{j,m}^{\mathbf{k}}} = z_{-j,m}^{\bar{\mathbf{k}}}$ and

$$z_{j,e(l)}^{\pm(l)} = 0 \quad \text{if} \quad |l| \neq |j|. \quad (38)$$

4.3 Bounds of the modulation functions

For polynomials $z = z(s)$, we introduce the norm

$$\|z\|_t = \sum_{l \geq 0} \frac{1}{l!} \left| \frac{d^l}{ds^l} z(s) \right|_{s=\delta^\nu t}$$

for $0 \leq t \leq 1$. It has the properties that $\|z \cdot w\|_t \leq \|z\|_t \cdot \|w\|_t$ and $\|\dot{v}\|_t \leq \deg(v) \|v\|_t$. With this norm, we have the following discrete counterpart of [15, Lemma 1]. In contrast to the analytical result in [15], we gain a factor $|\text{sinc}(\tau\omega_j)|^{-1}$ in the estimates, which comes from the filter ψ in the trigonometric integrator (6) and will be needed later on to deal with this factor in the almost-invariant energies \mathcal{E}_l of (23).

Lemma 10. *For $m = 1, \dots, 2K - 1$ and $0 \leq t \leq 1$ we have*

$$\begin{aligned} \sum_{j=-M}^{M-1} \sigma_j \left(|\text{sinc}(\tau\omega_j)|^{-1} \sum_{\pm\langle j \rangle \neq \mathbf{k} \in \mathcal{K}_j} \gamma_j^{\mathbf{k}} \|z_{j,m}^{\mathbf{k}}\|_t \right)^2 &\lesssim 1, \\ \sum_{j=-M}^{M-1} \sigma_j \left(|\text{sinc}(\tau\omega_j)|^{-1} \gamma_j^{\pm\langle j \rangle} \|z_{j,m}^{\pm\langle j \rangle}\|_t \right)^2 &\lesssim 1, \quad \sum_{j=-M}^{M-1} \sigma_j \left(\gamma_j^{\pm\langle j \rangle} \|z_{j,m}^{\pm\langle j \rangle}\|_t \right)^2 \lesssim 1 \end{aligned}$$

with the additional weight

$$\gamma_j^{\mathbf{k}} = \max(1, \omega_j, |\mathbf{k} \cdot \boldsymbol{\omega}|).$$

Proof. The statement is shown by induction on m , the case $m = 0$ being clear by notation ($z_{j,0}^{\mathbf{k}} = 0$).

(a) We first consider the case $\mathbf{k} \neq \pm\langle j \rangle$. Within this case, we first consider the case $|j| \leq K$, in which the strong non-resonance condition (12b) holds. In the notation (31) of (29) we thus have $1/|\alpha_0| \leq \gamma^{-2}\tau^{-2}$, $|\alpha_1|/|\alpha_0| \leq \gamma^{-1}\delta^\nu$ since $|s_{2\mathbf{k}}| \leq |s_{\langle j \rangle + \mathbf{k}}| + |s_{\langle j \rangle - \mathbf{k}}|$ and $|\alpha_l|/|\alpha_0| \leq \gamma^{-2}\delta^{2\nu}$ for $l \geq 2$. Using the symplecticity (7) to replace ψ by $\text{sinc} \cdot \phi$ in (29), the boundedness of ϕ , $m(j_1, \mathbf{k}^1) + m(j_2, \mathbf{k}^2) \geq m(j, \mathbf{k})$ and the properties $\|z \cdot w\|_t \leq \|z\|_t \cdot \|w\|_t$ and $\|\dot{p}\|_t \leq \deg(p) \|p\|_t$, we get

$$\|z_{j,m}^{\mathbf{k}}\|_t \lesssim |\text{sinc}(\tau\omega_j)| \sum_{j_1+j_2 \equiv j} \sum_{\mathbf{k}^1+\mathbf{k}^2=\mathbf{k}} \sum_{m_1+m_2=m} \|z_{j_1,m_1}^{\mathbf{k}^1}\|_t \|z_{j_2,m_2}^{\mathbf{k}^2}\|_t$$

for the solution $z = z_{j,m}^{\mathbf{k}}$ of (31) given by (32).

The same estimate also holds in the case $\mathbf{k} \neq \pm\langle j \rangle$ if $|j| > K$. This can be seen as follows. For these indices, only the weaker non-resonance condition (12a) holds, and hence we only have $1/|\alpha_0| \leq \gamma^{-2}\tau^{-2}\delta^{-2\nu}$, $|\alpha_1|/|\alpha_0| \leq \gamma^{-1}$ and $|\alpha_l|/|\alpha_0| \leq \gamma^{-2}$ for $l \geq 2$. The problematic factor $\delta^{-2\nu}$, however, can be compensated with the power of δ in the polynomial p on the right-hand side of (31). In fact, we have in this case

$$m(j, \mathbf{k}) = e(j) = K < e(j_1) + e(j_2) \leq m(j_1, \mathbf{k}^1) + m(j_2, \mathbf{k}^2)$$

since $|j| > K$, and hence there is an additional factor $\delta^{2\nu}$ in the polynomial p .

This shows that off-diagonal modulation coefficient functions $z_{j,m}^{\mathbf{k}}$ are bounded by a constant, since only modulation coefficient functions $z_{j',m'}^{\mathbf{k}'}$ with $m' < m$ can appear

on right-hand side of the above estimate for $z_{j,m}^{\mathbf{k}}$. To get the summed estimate of the lemma, we use $\gamma_j^{\mathbf{k}} \lesssim \gamma_{j_1}^{\mathbf{k}^1} \gamma_{j_2}^{\mathbf{k}^2}$ for $\mathbf{k} = \mathbf{k}^1 + \mathbf{k}^2$ and $j \equiv j_1 + j_2$, which yields

$$\sum_{j=-M}^{M-1} \sigma_j \left(|\text{sinc}(\tau\omega_j)|^{-1} \sum_{\pm\langle j \rangle \neq \mathbf{k} \in \mathcal{K}_j} \gamma_j^{\mathbf{k}} \|z_{j,m}^{\mathbf{k}}\|_t \right)^2 \lesssim \sum_{j=-M}^{M-1} \sigma_j \left(\sum_{j_1+j_2 \equiv j} v_{j_1} v_{j_2} \right)^2$$

with

$$v_{j_1} = \sum_{\mathbf{k}^1} \sum_{m_1=1}^{m-1} \gamma_{j_1}^{\mathbf{k}^1} \|z_{j_1,m_1}^{\mathbf{k}^1}\|_t, \quad v_{j_2} = \sum_{\mathbf{k}^2} \sum_{m_2=1}^{m-1} \gamma_{j_2}^{\mathbf{k}^2} \|z_{j_2,m_2}^{\mathbf{k}^2}\|_t.$$

The first estimate then follows inductively from the algebra property (21) and $1 \leq |\text{sinc}(\tau\omega_j)|^{-1}$.

(b) In the case $\mathbf{k} = \pm\langle j \rangle$, the absolute value of the coefficient α_1 is bounded from below by $\gamma\tau^2\delta^{2\nu}$ by the (weaker) non-resonance condition (12a). Also in this case, we have an additional factor $\delta^{2\nu}$ in the polynomial on the right-hand side of (31) since $m(j, \mathbf{k}) = e(j)$ and $m(j_1, \mathbf{k}^1) \geq e(j)$ or $m(j_2, \mathbf{k}^2) \geq e(j)$ for $\mathbf{k} = \pm\langle j \rangle = \mathbf{k}^1 + \mathbf{k}^2$. This shows that

$$\|z_{j,m}^{\pm\langle j \rangle}\|_t \lesssim |\text{sinc}(\tau\omega_j)| \sum_{j_1+j_2 \equiv j} \sum_{\mathbf{k}^1+\mathbf{k}^2=\pm\langle j \rangle} \sum_{m_1+m_2=m} \|z_{j_1,m_1}^{\mathbf{k}^1}\|_t \|z_{j_2,m_2}^{\mathbf{k}^2}\|_t$$

for the solution $z = z_{j,m}^{\pm\langle j \rangle}$ of (31) given by (33). As in (a), this yields the second estimate of the lemma.

(c) With the definitions of $\gamma_j^{\mathbf{k}}$ and of the norm $\|\cdot\|_t$ for $t = 0$ and with the estimates $1 \leq |\text{sinc}(\tau\omega_j)|^{-1}$ and $|\text{sinc}(\tau(\mathbf{k} \cdot \boldsymbol{\omega}))| \leq 1$, we get for the initial values $z_{j,m}^{\pm\langle j \rangle}(0)$ constructed with (34) that

$$\begin{aligned} \gamma_j^{\pm\langle j \rangle} |z_{j,m}^{\pm\langle j \rangle}(0)| &\lesssim \sum_{\substack{\mathbf{k} \in \mathcal{K}_j \\ \mathbf{k} \neq \langle j \rangle, \mathbf{k} \neq -\langle j \rangle}} |\text{sinc}(\tau\omega_j)|^{-1} \gamma_j^{\mathbf{k}} |z_{j,m}^{\mathbf{k}}(0)| \\ &\quad + \sum_{\mathbf{k} \in \mathcal{K}_j} |\text{sinc}(\tau\omega_j)|^{-1} \gamma_j^{\mathbf{k}} \|z_{j,m}^{\mathbf{k}}\|_0 + \delta^{-e(j)} \omega_j |u_j^0| + \delta^{-e(j)} |\dot{u}_j^0|. \end{aligned}$$

With the results of (a) and (b) and with the assumption (22) on the initial values, this yields

$$\sum_{j=-M}^{M-1} \sigma_j \left(\gamma_j^{\pm\langle j \rangle} |z_{j,m}^{\pm\langle j \rangle}(0)| \right)^2 \lesssim 1.$$

The last estimate of the lemma is a combination of this estimate and the second estimate of the lemma proven in (b). \square

By means of the expansion (28) of the modulation functions $z_j^{\mathbf{k}}$, the bounds of Lemma 10 on the coefficients $z_{j,m}^{\mathbf{k}}$ of this expansion imply in particular the bounds

$$\sum_{j=-M}^{M-1} \sigma_j \left(\sum_{\mathbf{k} \in \mathcal{K}_j} \gamma_j^{\mathbf{k}} \delta^{-m(j,\mathbf{k})} \|z_j^{\mathbf{k}}\|_t \right)^2 \lesssim 1, \quad (39)$$

$$\sum_{j=-M}^{M-1} \sigma_j \left(|\text{sinc}(\tau\omega_j)|^{-1} \sum_{\pm\langle j \rangle \neq \mathbf{k} \in \mathcal{K}_j} \gamma_j^{\mathbf{k}} \delta^{-m(j,\mathbf{k})} \|z_j^{\mathbf{k}}\|_t \right)^2 \lesssim 1, \quad (40)$$

on the modulation functions $z_j^{\mathbf{k}}$ themselves. In particular, we have

$$z_j^{\mathbf{k}} = \mathcal{O}(\delta^{e(j)}) \quad \text{and} \quad z_j^{\mathbf{k}} = \mathcal{O}(\delta^{e(l)}) \quad \text{if} \quad k_l \neq 0 \quad (41)$$

by the definition of $m(j, \mathbf{k})$.

4.4 Proof of Theorem 4

The proof of Theorem 4 relies on the alternative form (26) of the almost-invariant energies $\mathcal{E}_l(t)$ of (23). Note again that only modulation functions $z_j^{\mathbf{k}}$ with $k_l \neq 0$ can appear in $\mathcal{E}_l(t)$ due to the factor k_l . By (41), this argument shows that $\mathcal{E}_l(t) = \mathcal{O}(\delta^{2e(l)})$ has the claimed order in δ .

To get the precise estimate of Theorem 4, we multiply (26) with $\sigma_l \delta^{-2e(l)}$, we use $e(l) \leq m(j, \mathbf{k})$ for $k_l \neq 0$, and we sum over l , which yields

$$\begin{aligned} & \sum_{l=0}^M \sigma_l \delta^{-2e(l)} |\mathcal{E}_l(t)| \\ & \leq \frac{1}{2} \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \left(\sum_{l=0}^M \sigma_l |k_l| \omega_l \right) |\mathbf{k} \cdot \boldsymbol{\omega}| \left| \frac{\text{sinc}(\tau(\mathbf{k} \cdot \boldsymbol{\omega}))}{\text{sinc}(\tau \omega_j)} \right| \delta^{-2m(j, \mathbf{k})} |z_j^{\mathbf{k}}(\epsilon^{\nu/2} t)|^2 \\ & \quad + \frac{1}{2} \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \left(\sum_{l=0}^M \sigma_l |k_l| \omega_l \right) \frac{\delta^\nu}{|\text{sinc}(\tau \omega_j)|} \delta^{-2m(j, \mathbf{k})} |z_{-j}^{-\mathbf{k}}(\epsilon^{\nu/2} t)| \cdot \|z_j^{\mathbf{k}}\|_t \end{aligned}$$

by the definition of the norm $\|\cdot\|_t$. For the first term on the right-hand side of this estimate, we use

$$\sum_{l=0}^M \sigma_l |k_l| \omega_l \lesssim \sigma_j \gamma_j^{\mathbf{k}} \quad \text{if} \quad \mathbf{k} \in \mathcal{K}_j \quad (42)$$

and $|\mathbf{k} \cdot \boldsymbol{\omega}| \leq \gamma_j^{\mathbf{k}}$, we use that the ℓ^1 -norm dominates the ℓ^2 -norm (more precisely, $\sum_{\mathbf{k} \in \mathcal{K}_j} |v_j^{\mathbf{k}}|^2 \leq (\sum_{\mathbf{k} \in \mathcal{K}_j} |v_j^{\mathbf{k}}|)^2$), and we use (39) for $\mathbf{k} = \pm \langle j \rangle$ (the two sines cancel in this case) and (40) for $\mathbf{k} \neq \pm \langle j \rangle$. This shows that this first term is bounded by a constant. To show that also the second term is bounded by a constant, we apply in addition the non-resonance condition (12a) to bound $\delta^\nu / |\text{sinc}(\tau \omega_j)|$ and the Cauchy-Schwarz inequality.

4.5 Bounds of the defect

When constructing $z_j^{\mathbf{k}}$ with the expansion (28), the defect $d_j^{\mathbf{k}}$, $\mathbf{k} \in \mathcal{K}_j$, in (15) is given by

$$\begin{aligned} d_j^{\mathbf{k}} = & - \sum_{m=2K}^{2(2K-1)} \delta^{m(1-2\nu)} \sum_{\mathbf{k}^1 + \mathbf{k}^2 = \mathbf{k}} \sum_{j_1 + j_2 \equiv j} \delta^{2(m(j_1, \mathbf{k}^1) + m(j_2, \mathbf{k}^2))\nu} \\ & \sum_{m_1 + m_2 = m} \phi(\tau \omega_{j_1}) z_{j_1, m_1}^{\mathbf{k}^1} \phi(\tau \omega_{j_2}) z_{j_2, m_2}^{\mathbf{k}^2}. \end{aligned} \quad (43)$$

Although we consider system (15) defining the defect $d_j^{\mathbf{k}}$ only for $\mathbf{k} \in \mathcal{K}_j$, we use this formula to define $d_j^{\mathbf{k}}$ also for $\mathbf{k} \notin \mathcal{K}_j$. This will be helpful below.

Lemma 11. *The defect $d_j^{\mathbf{k}}$ in (15) given by (43) satisfies, for $0 \leq t \leq 1$,*

$$\sum_{j=-M}^{M-1} \sigma_j \left(\sum_{\mathbf{k} \in \mathbb{Z}^{M+1}} \gamma_j^{\mathbf{k}} \delta^{-2m(j, \mathbf{k})\nu} \|d_j^{\mathbf{k}}\|_t \right)^2 \lesssim \delta^{4K(1-2\nu)}$$

with $\gamma_j^{\mathbf{k}}$ as in Lemma 10.

Proof. The result follows with Lemma 10 and the arguments used in its proof. \square

On a higher level, the approximation $\tilde{\mathbf{u}}^n = (\tilde{u}_{-M}^n, \dots, \tilde{u}_{M-1}^n)^T$ of (14) to the numerical solution \mathbf{u}^n then has a defect $\mathbf{e}^n = (e_{-M}^n, \dots, e_{M-1}^n)^T$ when inserted into the numerical method (6a):

$$\tilde{\mathbf{u}}^{n+1} - 2 \cos(\tau \Omega) \tilde{\mathbf{u}}^n + \tilde{\mathbf{u}}^{n-1} = \tau^2 \Psi((\Phi \tilde{\mathbf{u}}^n) * (\Phi \tilde{\mathbf{u}}^n)) + \tau^2 \Psi \mathbf{e}^n. \quad (44)$$

By construction, this defect is given by

$$e_j^n = \sum_{\mathbf{k} \in \mathcal{K}_j} d_j^{\mathbf{k}} e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n} - \sum_{\mathbf{k} \in \mathbb{Z}^{M+1} \setminus \mathcal{K}_j} \sum_{\mathbf{k}^1 + \mathbf{k}^2 = \mathbf{k}} \sum_{j_1 + j_2 \equiv j} \phi(\tau \omega_{j_1}) z_{j_1}^{\mathbf{k}^1} \phi(\tau \omega_{j_2}) z_{j_2}^{\mathbf{k}^2} e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n},$$

where the defect $d_j^{\mathbf{k}}$ and the modulation functions are evaluated at $\delta^\nu t_n$.

Lemma 12. *The defect \mathbf{e}^n in (44) satisfies, for $0 \leq t_n = n\tau \leq 1$,*

$$\|\Omega \mathbf{e}^n\| \lesssim \delta^{2K(1-2\nu)}$$

with the norm $\|\cdot\|$ of Section 3.2.

Proof. We use that the set \mathcal{K}_j was constructed in such a way that \mathbf{k} belongs to \mathcal{K}_j if $\mathbf{k} = \mathbf{k}^1 + \mathbf{k}^2$, $j \equiv j_1 + j_2$ and $m_1 + m_2 \leq 2K - 1$ for $\mathbf{k}^1, \mathbf{k}^2, j_1, j_2, m_1, m_2$ with $z_{j_1, m_1}^{\mathbf{k}^1} \neq 0$ and $z_{j_2, m_2}^{\mathbf{k}^2} \neq 0$. Using (43) to define $d_j^{\mathbf{k}}$ also for $\mathbf{k} \notin \mathcal{K}_j$, this yields the compact form

$$e_j^n = \sum_{\mathbf{k} \in \mathbb{Z}^{M+1}} d_j^{\mathbf{k}} (\delta^\nu t_n) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n}$$

for the defect in (44). From the definition of the norm $\|\cdot\|$ (see Section 3.2), we thus get

$$\|\Omega \mathbf{e}^n\|^2 \leq \sum_{j=-M}^{M-1} \sigma_j \delta^{-4e(j)\nu} \left(\sum_{\mathbf{k} \in \mathbb{Z}^{M+1}} \omega_j |d_j^{\mathbf{k}}(\delta^\nu t_n)| \right)^2.$$

Using $e(j) \leq m(j, \mathbf{k})$ and $\omega_j \leq \gamma_j^{\mathbf{k}}$, the statement of the lemma thus follows from Lemma 11. \square

4.6 Bounds of the remainder

With the constructed modulation functions $z_j^{\mathbf{k}}$, we get the approximation \tilde{u}_j^n of (14) to the numerical solution u_j^n . From (17), we also get an approximation \hat{u}_j^n to \dot{u}_j^n . In the following lemma, we establish a bound for the approximation error.

Lemma 13. *The remainders $u_j^n - \tilde{u}_j^n$ and $\dot{u}_j^n - \hat{u}_j^n$ satisfy, for $0 \leq t_n \leq 1$,*

$$\|\Omega(\mathbf{u}^n - \tilde{\mathbf{u}}^n)\| + \|\dot{\mathbf{u}}^n - \hat{\dot{\mathbf{u}}}^n\| \lesssim \delta^{2K(1-2\nu)}.$$

Proof. (a) In a first step, we write the numerical scheme (6) and its approximation by a modulated Fourier expansion in one-step form. The method in one-step form reads

$$\begin{pmatrix} \mathbf{u}^{n+1} \\ \dot{\mathbf{u}}^{n+1} \end{pmatrix} = \begin{pmatrix} \cos(\tau\Omega) & \Omega^{-1} \sin(\tau\Omega) \\ -\Omega \sin(\tau\Omega) & \cos(\tau\Omega) \end{pmatrix} \begin{pmatrix} \mathbf{u}^n \\ \dot{\mathbf{u}}^n \end{pmatrix} + \frac{\tau}{2} \begin{pmatrix} \tau\Psi\mathbf{g}^n \\ \cos(\tau\Omega)\Phi\mathbf{g}^n + \Phi\mathbf{g}^{n+1} \end{pmatrix} \quad (45)$$

with $\mathbf{g}^n = (\Phi\mathbf{u}^n) * (\Phi\mathbf{u}^n)$, see [18, Section XIII.2.2]. The first line is obtained by adding (6a) and (6c), and the second line is obtained by subtracting these equations with $n+1$ instead of n and using the first line to replace \mathbf{u}^{n+1} as well as the symplecticity (7). In the same way, we derive for the approximations $\tilde{\mathbf{u}}^n$ and $\dot{\tilde{\mathbf{u}}}^n$, which satisfy (6c) exactly (by definition (17)) and (6a) up to a small defect given by (44),

$$\begin{pmatrix} \tilde{\mathbf{u}}^{n+1} \\ \dot{\tilde{\mathbf{u}}}^{n+1} \end{pmatrix} = \begin{pmatrix} \cos(\tau\Omega) & \Omega^{-1} \sin(\tau\Omega) \\ -\Omega \sin(\tau\Omega) & \cos(\tau\Omega) \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{u}}^n \\ \dot{\tilde{\mathbf{u}}}^n \end{pmatrix} + \frac{\tau}{2} \begin{pmatrix} \tau\Psi\tilde{\mathbf{g}}^n \\ \cos(\tau\Omega)\Phi\tilde{\mathbf{g}}^n + \Phi\tilde{\mathbf{g}}^{n+1} \end{pmatrix} \quad (46)$$

with $\tilde{\mathbf{g}}^n = \mathbf{g}^n + \mathbf{e}^n$.

(b) In the following, we write

$$\|(\mathbf{v}, \dot{\mathbf{v}})\| = (\|\Omega\mathbf{v}\|^2 + \|\dot{\mathbf{v}}\|^2)^{1/2} \quad (47)$$

for a norm that is equivalent to the norm considered in the statement of the lemma. By induction on n , we prove that the numerical solutions stays small in this norm,

$$\|(\mathbf{u}^n, \dot{\mathbf{u}}^n)\| \leq (\sqrt{2C_0} + t_n)\delta^{1-2\nu} \quad \text{for} \quad 0 \leq t_n \leq 1$$

with C_0 from (22), provided that $\delta^{1-2\nu}$ is sufficiently small. This is true for $n=0$ by the choice of the initial value (22). For $n>0$, we first note that, by (45),

$$\|\Omega\mathbf{u}^n\| \leq \|\Omega\mathbf{u}^{n-1}\| + \|\dot{\mathbf{u}}^{n-1}\| + \frac{1}{2}\tau^2\|\Omega\Psi\mathbf{g}^{n-1}\|,$$

and hence $\|\Omega\mathbf{u}^n\| \leq C\delta^{1-2\nu}$ by induction, by (7), by the boundedness of ϕ and by the algebra property (20). We then observe that the matrix appearing in the one-step formulation (45) of the method preserves the norm (47). Moreover, by the algebra properties (19) and (20), the second term in the one-step formulation (45) can be estimated in the norm (47) by $\tau\delta^{2(1-2\nu)}$ up to a constant. This implies the stated bound of $\|(\mathbf{u}^n, \dot{\mathbf{u}}^n)\|$.

Similarly, we get for the modulated Fourier expansion the bound

$$\|(\tilde{\mathbf{u}}^n, \dot{\tilde{\mathbf{u}}}^n)\| \leq (\sqrt{2C_0} + t_n)\delta^{1-2\nu} \quad \text{for} \quad 0 \leq t_n \leq 1.$$

To get this bound, we just repeat the presented argument for the numerical solution given by (45) for the modulated Fourier expansion which satisfies the very similar relation (46). The only difference is the additional term \mathbf{e}^n in (46), which can be estimated with Lemma 12.

(c) For the difference $(\mathbf{u}^n - \tilde{\mathbf{u}}^n, \dot{\mathbf{u}}^n - \dot{\tilde{\mathbf{u}}}^n)$, we subtract the above one-step formulations, and then proceed as in the proof of the smallness of $(\mathbf{u}^n, \dot{\mathbf{u}}^n)$ and $(\tilde{\mathbf{u}}^n, \dot{\tilde{\mathbf{u}}}^n)$ in (b). We use the smallness of $(\mathbf{u}^n, \dot{\mathbf{u}}^n)$ and $(\tilde{\mathbf{u}}^n, \dot{\tilde{\mathbf{u}}}^n)$ together with the algebra property (19) and Lemma 12 to control the difference $\mathbf{g}^n - \tilde{\mathbf{g}}^n = -\mathbf{e}^n$. The mentioned lemma on the defect introduces the small parameter $\delta^{2K(1-2\nu)}$. \square

With the results proven in this section, the proofs of Theorems 3 and 4 are complete.

5 Almost-invariant energies: Proofs of Theorems 5–7

5.1 Almost-invariant energies: Proof of Theorem 5

For the proof of Theorem 5, we sum the equality (24) to get

$$\mathcal{E}_l(t_n) - \mathcal{E}_l(0) = -\frac{i}{2} \sum_{j=-M}^{M-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \sum_{\tilde{n}=1}^n \tau k_l \omega_l \phi(\tau \omega_j) z_{-j}^{-\mathbf{k}}(\delta^\nu t_{\tilde{n}}) d_j^{\mathbf{k}}(\delta^\nu t_{\tilde{n}})$$

for the l th almost-invariant energy \mathcal{E}_l . Note again that only indices \mathbf{k} with $k_l \neq 0$ are relevant in the above sum, and hence we have there $z_{-j}^{-\mathbf{k}} = \mathcal{O}(\delta^{e(l)})$ by (41) and $d_j^{\mathbf{k}} = \mathcal{O}(\delta^{e(l)+(2K-e(l))(1-2\nu)})$ by Lemma 11 (since $m(j, \mathbf{k}) \geq e(l)$). This shows that the variation in the almost-invariant energy \mathcal{E}_l is indeed of the claimed order, and it shows the additional statement for $l = 1$.

To get the precise estimate of Theorem 5, we proceed as in the proof of Theorem 4 in Section 4.4: We multiply the above equation with $\sigma_l \delta^{-2e(l)}$ and sum over l , we use $e(l) \leq m(j, \mathbf{k}) \leq K$ for $k_l \neq 0$, (42) and $n\tau = t_n \leq 1$, and we apply the Cauchy–Schwarz inequality, the estimate (39) of the modulation functions, and the estimate of the defect of Lemma 11.

5.2 Transitions in the almost-invariant energies: Proof of Theorem 6

We consider the situation of Theorem 6, with a modulated Fourier expansion with coefficients $z_j^{\mathbf{k}}(\delta^\nu t)$, $0 \leq t \leq 1$, constructed from $(\mathbf{u}^0, \dot{\mathbf{u}}^0)$ and a modulated Fourier expansion with coefficients $\tilde{z}_j^{\mathbf{k}}(\delta^\nu t)$, $0 \leq t \leq 1$, constructed from $(\mathbf{u}^N, \dot{\mathbf{u}}^N)$, where N is such that $t_N = N\tau = 1$. For studying the difference of the corresponding almost-invariant energies, we first consider the difference of the modulation functions themselves.

Lemma 14. *Under the assumptions of Theorem 6, we have*

$$\sum_{j=-M}^{M-1} \sigma_j \left(\sum_{\mathbf{k} \in \mathcal{K}_j} \gamma_j^{\mathbf{k}} \delta^{-2m(j, \mathbf{k})\nu} \left\| z_j^{\mathbf{k}}(\cdot + \delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} - \tilde{z}_j^{\mathbf{k}} \right\|_0 \right)^2 \lesssim \delta^{4K(1-2\nu)}.$$

Proof. Recall that the functions $z_j^{\mathbf{k}}$ and $\tilde{z}_j^{\mathbf{k}}$ are constructed with the expansion (28). We consider here the truncated expansions

$$\begin{aligned} [z_j^{\mathbf{k}}]^\ell(\delta^\nu t) &= \delta^{m(j, \mathbf{k})} \sum_{m=m(j, \mathbf{k})}^{\ell} \delta^{(m-m(j, \mathbf{k}))(1-2\nu)} z_{j, m}^{\mathbf{k}}(\delta^\nu t), \\ [\tilde{z}_j^{\mathbf{k}}]^\ell(\delta^\nu t) &= \delta^{m(j, \mathbf{k})} \sum_{m=m(j, \mathbf{k})}^{\ell} \delta^{(m-m(j, \mathbf{k}))(1-2\nu)} \tilde{z}_{j, m}^{\mathbf{k}}(\delta^\nu t) \end{aligned}$$

for $\ell = 1, \dots, 2K-1$. Note that these truncations coincide for $\ell = 2K-1$ with $z_j^{\mathbf{k}}$

and $\tilde{z}_j^{\mathbf{k}}$, respectively. We therefore study the differences

$$\begin{aligned} [f_j^{\mathbf{k}}]^\ell(\delta^\nu t) &= \delta^{-2m(j,\mathbf{k})\nu} \left([z_j^{\mathbf{k}}]^\ell(\delta^\nu t + \delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} - [\tilde{z}_j^{\mathbf{k}}]^\ell(\delta^\nu t) \right) \\ &= \sum_{m=m(j,\mathbf{k})}^{\ell} \delta^{m(1-2\nu)} (z_{j,m}^{\mathbf{k}}(\delta^\nu t + \delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} - \tilde{z}_{j,m}^{\mathbf{k}}(\delta^\nu t)). \end{aligned}$$

(a) We first derive equations for these differences from the defining equations (29) and (34). For $\mathbf{k} \neq \pm \langle j \rangle$ we multiply equation (29) by $\delta^{m(1-2\nu)}$ and sum over m for $z_{j,m}^{\mathbf{k}}$ and its derivatives evaluated at $\delta^\nu t + \delta^\nu$ and in the same way for $\tilde{z}_{j,m}^{\mathbf{k}}$ at $\delta^\nu t$. By multiplying the first sum with $e^{i(\mathbf{k} \cdot \boldsymbol{\omega})}$, we get for the difference of these sums

$$\begin{aligned} &4s_{\langle j \rangle - \mathbf{k}} s_{\langle j \rangle + \mathbf{k}} [f_j^{\mathbf{k}}]^\ell + 2i\tau\delta^\nu s_{2\mathbf{k}} [\dot{f}_j^{\mathbf{k}}]^\ell + \tau^2\delta^{2\nu} c_{2\mathbf{k}} [\ddot{f}_j^{\mathbf{k}}]^\ell + \dots \\ &= \tau^2\psi(\tau\omega_j) \sum_{\mathbf{k}^1 + \mathbf{k}^2 = \mathbf{k}} \sum_{j_1 + j_2 \equiv j} \sum_{m=m(j,\mathbf{k})}^{\ell} \sum_{m_1 + m_2 = m} \delta^{m(1-2\nu) + 2(m(j_1, \mathbf{k}^1) + m(j_2, \mathbf{k}^2) - m(j, \mathbf{k}))\nu} \\ &\quad \phi(\tau\omega_{j_1}) \phi(\tau\omega_{j_2}) \left(z_{j_1, m_1}^{\mathbf{k}^1} e^{i(\mathbf{k}^1 \cdot \boldsymbol{\omega})} z_{j_2, m_2}^{\mathbf{k}^2} e^{i(\mathbf{k}^2 \cdot \boldsymbol{\omega})} - \tilde{z}_{j_1, m_1}^{\mathbf{k}^1} \tilde{z}_{j_2, m_2}^{\mathbf{k}^2} \right), \end{aligned}$$

where $\tilde{z}_{j_i, m_i}^{\mathbf{k}^i}$ and $f_j^{\mathbf{k}}$ are evaluated at $\delta^\nu t$ and $z_{j_i, m_i}^{\mathbf{k}^i}$ are evaluated at $\delta^\nu t + \delta^\nu$ for $i = 1, 2$. We use $a_1 a_2 - \tilde{a}_1 \tilde{a}_2 = (a_1 - \tilde{a}_1) a_2 + \tilde{a}_1 (a_2 - \tilde{a}_2)$, the symmetry of the sums in j_l and \mathbf{k}^l , the convention $z_{j, m'}^{\mathbf{k}} = 0$ and $\tilde{z}_{j, m'}^{\mathbf{k}} = 0$ for $m' < m(j, \mathbf{k})$, and $\sum_{m=1}^{\ell} \sum_{m_1 + m_2 = m} \% = \sum_{m_2=1}^{\ell-1} \sum_{m_1=1}^{\ell-m_2} \%$. This yields

$$\begin{aligned} &4s_{\langle j \rangle - \mathbf{k}} s_{\langle j \rangle + \mathbf{k}} [f_j^{\mathbf{k}}]^\ell + 2i\tau\delta^\nu s_{2\mathbf{k}} [\dot{f}_j^{\mathbf{k}}]^\ell + \tau^2\delta^{2\nu} c_{2\mathbf{k}} [\ddot{f}_j^{\mathbf{k}}]^\ell + \dots \\ &= \tau^2\psi(\tau\omega_j) \sum_{\mathbf{k}^1 + \mathbf{k}^2 = \mathbf{k}} \sum_{j_1 + j_2 \equiv j} \delta^{2(m(j_1, \mathbf{k}^1) + m(j_2, \mathbf{k}^2) - m(j, \mathbf{k}))\nu} \\ &\quad \sum_{m_2=m(j_2, \mathbf{k}^2)}^{\ell-1} \delta^{m_2(1-2\nu)} \phi(\tau\omega_{j_1}) [f_{j_1}^{\mathbf{k}^1}]^{\ell-m_2} \phi(\tau\omega_{j_2}) \left(\tilde{z}_{j_2, m_2}^{\mathbf{k}^2} + z_{j_2, m_2}^{\mathbf{k}^2} e^{i(\mathbf{k}^2 \cdot \boldsymbol{\omega})} \right). \end{aligned}$$

Using the truncated expansion $[z_j^{\mathbf{k}}]^\ell$ instead of ansatz (28) we get as in Section 4 corresponding to Theorem 3

$$\begin{aligned} u_j^N &= \sum_{\mathbf{k} \in \mathcal{K}_j} [z_j^{\mathbf{k}}]^\ell(\delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} + [r_j^N]^\ell \\ \dot{u}_j^N &= (\tau \operatorname{sinc}(\tau\omega_j))^{-1} \sum_{\mathbf{k} \in \mathcal{K}_j} \left(i s_{2\mathbf{k}} [z_j^{\mathbf{k}}]^\ell(\delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} + \tau \delta^\nu c_{2\mathbf{k}} [\dot{z}_j^{\mathbf{k}}]^\ell(\delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} \right. \\ &\quad \left. + \frac{i}{2} \tau^2 \delta^{2\nu} s_{2\mathbf{k}} [\ddot{z}_j^{\mathbf{k}}]^\ell(\delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} + \dots \right) + [\dot{r}_j^N]^\ell, \end{aligned}$$

where the remainders $[r_j^N]^\ell$ and $[\dot{r}_j^N]^\ell$ of the truncated expansion satisfy

$$\|\Omega[\mathbf{r}^N]^\ell\| + \|\dot{\mathbf{r}}^N\|^\ell \lesssim \delta^{(\ell+1)(1-2\nu)}.$$

Note that the bound in Theorem 3 in terms of δ equals $C\delta^{(2K-1+1)(1-2\nu)}$ for the expansion truncated at $2K-1$. Inserting this into (34) for \tilde{z} (with u_j^N and \dot{u}_j^N instead

of u_j^0 and \dot{u}_j^0) yields, for $\ell \geq e(j)$,

$$\begin{aligned} 2i\omega_j [f_j^{\pm\langle j \rangle}]^\ell &= -i \sum_{\substack{\mathbf{k} \in \mathcal{K}_j \\ \mathbf{k} \neq \langle j \rangle, \mathbf{k} \neq -\langle j \rangle}} \delta^{2(m(j, \mathbf{k}) - e(j))\nu} \left(\omega_j \pm (\mathbf{k} \cdot \boldsymbol{\omega}) \frac{\text{sinc}(\tau(\mathbf{k} \cdot \boldsymbol{\omega}))}{\text{sinc}(\tau\omega_j)} \right) [f_j^{\mathbf{k}}]^\ell \\ &\mp \sum_{\mathbf{k} \in \mathcal{K}_j} \frac{\delta^{2(m(j, \mathbf{k}) - e(j))\nu}}{\text{sinc}(\tau\omega_j)} \left(\delta^\nu c_{2\mathbf{k}} [f_j^{\mathbf{k}}]^\ell + \frac{i}{2} \tau \delta^{2\nu} s_{2\mathbf{k}} [f_j^{\mathbf{k}}]^\ell + \dots \right) \\ &+ \delta^{-2e(j)\nu} \left(i\omega_j [r_j^N]^\ell \pm [\dot{r}_j^N]^\ell \right), \end{aligned}$$

where $f_j^{\mathbf{k}}$ is evaluated at 0.

(b) Based on the equations derived in (a), which correspond to equations (29) and (34), it can then be shown as in the proof Lemma 10 by induction on ℓ that

$$\sum_{j=-M}^{M-1} \sigma_j \left(\sum_{\mathbf{k} \in \mathcal{K}_j} \gamma_j^{\mathbf{k}} \| [f_j^{\mathbf{k}}]^\ell \|_0 \right)^2 \lesssim \delta^{2(\ell+1)(1-2\nu)}.$$

For $\ell = 2K - 1$ this is the bound as stated in the lemma. \square

Now we deduce Theorem 6 from this lemma. We consider the difference $\mathcal{E}_l(1) - \tilde{\mathcal{E}}_l(0)$ of almost-invariants in the form (26). This difference consists of differences of the form

$$z_{-j}^{-\mathbf{k}}(\delta^\nu) (z_j^{\mathbf{k}})^{(p)}(\delta^\nu) - \tilde{z}_{-j}^{-\mathbf{k}}(0) (\tilde{z}_j^{\mathbf{k}})^{(p)}(0)$$

with $p = 0, 1, 2, \dots$ denoting derivatives. We rewrite these differences, using

$$a_1 a_2 - \tilde{a}_1 \tilde{a}_2 = (a_1 e^{-i(\mathbf{k} \cdot \boldsymbol{\omega})} - \tilde{a}_1) a_2 e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} + \tilde{a}_1 (a_2 e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} - \tilde{a}_2).$$

The statement of Theorem 6 is then obtained by inserting ansatz (28) for $z_j^{\mathbf{k}}$ and $\tilde{z}_j^{\mathbf{k}}$ in $\sum_{l=0}^M \sigma_l \delta^{-2e(l)} |\mathcal{E}_l(1) - \tilde{\mathcal{E}}_l(0)|$ and applying Lemmas 10 and 14 together with (42). Note that by the previous Lemma 14 we have $z_j^{\mathbf{k}}(\cdot + \delta^\nu) e^{i(\mathbf{k} \cdot \boldsymbol{\omega})} - \tilde{z}_j^{\mathbf{k}} = \mathcal{O}(\delta^{e(l) + (2K - e(l))(1-2\nu)})$ if $k_l \neq 0$, and by (41) we have $z_j^{\mathbf{k}} = \mathcal{O}(\delta^{e(l)})$ and $\tilde{z}_j^{\mathbf{k}} = \mathcal{O}(\delta^{e(l)})$ if $k_l = 0$. This yields the claimed order $\delta^{2e(l) + K(1-2\nu)}$ of the transition in the l th almost-invariant energy, i.e., of $\mathcal{E}_l(1) - \tilde{\mathcal{E}}_l(0)$, and it yields the additional statement for $l = 1$.

5.3 Controlling mode energies by almost-invariant energies: Proof of Theorem 7

The proof of Theorem 7 is done in three steps. We first show in Lemma 15 below that the goal of controlling mode energies can be achieved by controlling certain dominant terms in the modulated Fourier expansion. Then we show, in Lemma 16 below, that these dominant terms can be controlled if only the diagonal ones among them are under control. Finally, we show in Lemma 17 below that these diagonal dominant terms can be described essentially by the almost-invariant energies of the modulated Fourier expansion. Throughout, we assume the conditions of Theorem 7 to be fulfilled.

Lemma 15. *For $0 \leq t \leq 1$ assume that*

$$\begin{aligned} |\text{sinc}(\tau\omega_j)|^{-1} \gamma_j^{\mathbf{k}} |z_{j,e(j)}^{\mathbf{k}}(\delta^\nu t)| &\leq \mathcal{C}_0 \quad \text{for} \quad \mathbf{k} \neq \pm\langle j \rangle, |j| \leq K, \mu(\mathbf{k}) \leq K, \\ \sum_{j=-M}^{M-1} \sigma_j \omega_j^2 |z_{j,e(j)}^{\pm\langle j \rangle}(\delta^\nu t)|^2 &\leq \mathcal{C}_0. \end{aligned}$$

Then, for $0 \leq t_n = n\tau \leq 1$,

$$\sum_{l=0}^M \sigma_l \delta^{-2e(l)} E_l^n \leq \mathcal{C}$$

and

$$\sigma_1 \delta^{-2e(1)} \left| E_1^n - \omega_1^2 \left(|z_1^{(1)}(\delta^\nu t_n)|^2 + |z_1^{-(1)}(\delta^\nu t_n)|^2 \right) \right| \leq \mathcal{C} \delta^{2(1-2\nu)},$$

where \mathcal{C} depends on \mathcal{C}_0 , but is independent of C_0 of (22) if $\delta^{1-2\nu}$ is sufficiently small.

Proof. We consider the two terms $\frac{1}{2}\omega_l^2 |u_l^n|^2$ and $\frac{1}{2}|\dot{u}_l^n|^2$ in E_l^n separately.

(a) For the first term, we use $|u_l^n| \leq |u_l^n - \tilde{u}_l^n| + |\tilde{u}_l^n|$ to get

$$\frac{1}{2} \sum_{l=0}^M \sigma_l \omega_l^2 \delta^{-2e(l)} |u_l^n|^2 \leq \sum_{l=0}^M \sigma_l \omega_l^2 \delta^{-2e(l)} |u_l^n - \tilde{u}_l^n|^2 + \sum_{l=0}^M \sigma_l \omega_l^2 \delta^{-2e(l)} |\tilde{u}_l^n|^2,$$

where the first sum is bounded by $C\delta^{2K(1-2\nu)}$ corresponding to Lemma 13. With the notation \tilde{u} of (14) and the expansion (28) for $z_l^{\mathbf{k}}$ we find

$$|\tilde{u}_l^n| \leq \sum_{\mathbf{k} \in \mathcal{K}_l} \delta^{m(l, \mathbf{k})} \left| \sum_{m=m(l, \mathbf{k})}^{2K-1} \delta^{(m-m(l, \mathbf{k}))(1-2\nu)} z_{l, m}^{\mathbf{k}}(\delta^\nu t_n) \right|,$$

where the dominant terms are $\delta^{m(l, \mathbf{k})} z_{l, m(l, \mathbf{k})}^{\mathbf{k}}$. Note that $m(l, \mathbf{k}) \geq e(l)$. This yields

$$\sum_{l=0}^M \sigma_l \omega_l^2 \delta^{-2e(l)} |\tilde{u}_l^n|^2 \leq \mathcal{C} + C\delta^{2(1-2\nu)},$$

with constants \mathcal{C} and C . The constant \mathcal{C} is the result of estimating the dominant terms $z_{l, m(l, \mathbf{k})}^{\mathbf{k}}$ with $m(l, \mathbf{k}) = e(l)$ using the assumption on these terms (note that $z_{l, m(l, \mathbf{k})}^{\mathbf{k}} = 0$ for $|l| > K$ and $\mathbf{k} \neq \pm \langle l \rangle$ by (36) and (37) since $m(l, \mathbf{k}) = e(l) = K$ in this case). The constant \mathcal{C} thus depends only on \mathcal{C}_0 and not on C_0 of (22). The second term $C\delta^{2(1-2\nu)}$ is the result of estimating the terms $z_{l, m(l, \mathbf{k})}^{\mathbf{k}}$ with $m(l, \mathbf{k}) > e(l)$ using Lemma 10. The constant C thus depends on C_0 of (22).

(b) For the second term, we proceed similarly. Using $|\dot{u}_l^n| \leq |\dot{u}_l^n - \dot{\tilde{u}}_l^n| + |\dot{\tilde{u}}_l^n|$ and again Lemma 13 to control the difference $|\dot{u}_l^n - \dot{\tilde{u}}_l^n|$, we just have to estimate $\sum_{l=0}^M \sigma_l \delta^{-2e(l)} |\dot{\tilde{u}}_l^n|^2$. We start from

$$|\dot{\tilde{u}}_l^n| \leq \sum_{\mathbf{k} \in \mathcal{K}_l} |\mathbf{k} \cdot \boldsymbol{\omega}| \frac{|\text{sinc}(\tau(\mathbf{k} \cdot \boldsymbol{\omega}))|}{|\text{sinc}(\tau\omega_l)|} |z_l^{\mathbf{k}}(\delta^\nu t_n)| + |\text{sinc}(\tau\omega_l)|^{-1} \sum_{\mathbf{k} \in \mathcal{K}_l} \|\dot{z}_l^{\mathbf{k}}\|_{t_n},$$

which is obtained using the velocity approximation \tilde{u} given by (17) and a Taylor expansion of $z_l^{\mathbf{k}}(\epsilon^{\nu/2} t_{n \pm 1})$. We then insert the expansion (28) of $z_l^{\mathbf{k}}$ and isolate the dominant terms $\delta^{m(l, \mathbf{k})} z_{l, m(l, \mathbf{k})}^{\mathbf{k}}$. Using the assumption on the terms $z_{l, e(l)}^{\mathbf{k}}$ and noting that $\dot{z}_{l, e(l)}^{\mathbf{k}} = 0$ by (35), this yields similarly as in (a)

$$\sum_{l=0}^M \sigma_l \delta^{-2e(l)} |\dot{\tilde{u}}_l^n|^2 \leq \mathcal{C} + C\delta^{2(1-2\nu)}.$$

Combining this estimate with the corresponding estimate in (a), we get the estimate of the lemma, provided that $\delta^{1-2\nu}$ is sufficiently small.

(c) The additional estimate of the difference $E_1^n - \omega_1^2(|z_1^{\langle 1 \rangle}(\delta^\nu t_n)|^2 + |z_1^{-\langle 1 \rangle}(\delta^\nu t_n)|^2)$ is obtained as follows. We first note that, as in [15],

$$\begin{aligned} 2E_1^n &= |\omega_1 u_1^n|^2 + |\dot{u}_1^n|^2 \\ &= \omega_1^2 |z_1^{\langle 1 \rangle}(\delta^\nu t_n) e^{i\omega_1 t_n} + z_1^{-\langle 1 \rangle}(\delta^\nu t_n) e^{-i\omega_1 t_n}|^2 \\ &\quad + 2\omega_1^2 \operatorname{Re} \left(\overline{\eta^n} (z_1^{\langle 1 \rangle}(\delta^\nu t_n) e^{i\omega_1 t_n} + z_1^{-\langle 1 \rangle}(\delta^\nu t_n) e^{-i\omega_1 t_n}) \right) + \omega_1^2 |\eta^n|^2 \\ &\quad + \omega_1^2 |iz_1^{\langle 1 \rangle}(\delta^\nu t_n) e^{i\omega_1 t_n} - iz_1^{-\langle 1 \rangle}(\delta^\nu t_n) e^{-i\omega_1 t_n}|^2 \\ &\quad + 2\omega_1 \operatorname{Re} \left(\overline{\vartheta^n} (iz_1^{\langle 1 \rangle}(\delta^\nu t_n) e^{i\omega_1 t_n} - iz_1^{-\langle 1 \rangle}(\delta^\nu t_n) e^{-i\omega_1 t_n}) \right) + |\vartheta^n|^2, \end{aligned}$$

but now with

$$\begin{aligned} \eta^n &= \sum_{\pm \langle 1 \rangle \neq \mathbf{k} \in \mathcal{K}_1} z_1^{\mathbf{k}}(\delta^\nu t_n) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n} + (u_1^n - \tilde{u}_1^n) \\ \vartheta^n &= \sum_{\pm \langle 1 \rangle \neq \mathbf{k} \in \mathcal{K}_1} i(\mathbf{k} \cdot \boldsymbol{\omega}) \frac{\operatorname{sinc}(\tau(\mathbf{k} \cdot \boldsymbol{\omega}))}{\operatorname{sinc}(\tau\omega_1)} z_1^{\mathbf{k}}(\delta^\nu t_n) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n} + (\dot{u}_1^n - \dot{\tilde{u}}_1^n) \\ &\quad + (\operatorname{sinc}(\tau\omega_1))^{-1} \sum_{\mathbf{k} \in \mathcal{K}_1} (c_{2\mathbf{k}} \delta^\nu \dot{z}_1^{\mathbf{k}}(\delta^\nu t_n) + \dots) e^{i(\mathbf{k} \cdot \boldsymbol{\omega}) t_n}, \end{aligned}$$

which is again obtained using the velocity approximation \tilde{u} given by (17) for $j = 1$ and a Taylor expansion of $z_1^{\mathbf{k}}(\epsilon^{\nu/2} t_{n\pm 1})$. Inserting the expansion (28) for $j = 1$, we get $|\eta^n| \lesssim \delta^{1+2(1-2\nu)}$ and $|\vartheta^n| \lesssim \delta^{1+2(1-2\nu)}$ by Lemma 10 and Lemma 13, since $z_{1,1}^{\mathbf{k}} = z_{1,2}^{\mathbf{k}} = 0$ for $\mathbf{k} \neq \pm \langle 1 \rangle$ and $\dot{z}_{1,1}^{\mathbf{k}} = \dot{z}_{1,2}^{\mathbf{k}} = 0$ for all \mathbf{k} . This yields the claimed estimate. \square

Lemma 16. For $0 \leq t \leq 1$ assume that

$$|z_{j,e(j)}^{\pm \langle j \rangle}(\delta^\nu t)| \leq C_0 \quad \text{for} \quad |j| \leq K.$$

Then,

$$|\operatorname{sinc}(\tau\omega_j)|^{-1} \gamma_j^{\mathbf{k}} |z_{j,e(j)}^{\mathbf{k}}(\delta^\nu t)| \leq C \quad \text{for} \quad \mathbf{k} \neq \pm \langle j \rangle, |j| \leq K, \mu(\mathbf{k}) \leq K$$

where C depends on C_0 but is independent of C_0 of (22).

Proof. This follows from the construction of off-diagonal modulation functions as in the case of the exact solution in [15, Lemma 4], see also the proof of Lemma 10. \square

Lemma 17. For $0 \leq t \leq 1$ we have

$$\begin{aligned} \delta^{-2e(0)} \left| \mathcal{E}_0(t) - \frac{1}{2} \omega_0^2 \left(|z_0^{\langle 0 \rangle}(\delta^\nu t)|^2 + |z_0^{-\langle 0 \rangle}(\delta^\nu t)|^2 \right) \right| &\lesssim \delta^{1-2\nu} \\ \delta^{-2e(1)} \left| \mathcal{E}_1(t) - \omega_1^2 \left(|z_1^{\langle 1 \rangle}(\delta^\nu t)|^2 + |z_1^{-\langle 1 \rangle}(\delta^\nu t)|^2 \right) \right| &\lesssim \delta^{2(1-2\nu)} \\ \sum_{l=1}^M \sigma_l \delta^{-2e(l)} \left| \mathcal{E}_l(t) - \omega_l^2 \left(|z_l^{\langle l \rangle}(\delta^\nu t)|^2 + |z_l^{-\langle l \rangle}(\delta^\nu t)|^2 \right) \right| &\lesssim \delta^{1-2\nu}. \end{aligned}$$

Proof. We subtract $\omega_l^2(|z_l^{\langle l \rangle}|^2 + |z_l^{-\langle l \rangle}|^2)$ for $l \geq 1$ and $\frac{1}{2}\omega_0^2(|z_0^{\langle 0 \rangle}|^2 + |z_0^{-\langle 0 \rangle}|^2)$ for $l = 0$ from the almost-invariant energy $\mathcal{E}_l(t)$ in the form (26). For the modulation functions that appear in this difference we still have $k_l \neq 0$ and consequently $m(j, \mathbf{k}) \geq e(l)$ and $k \neq \pm \langle j \rangle$ for $|j| \neq l$. By the expansion (28) and Lemma 10 we get for these functions

- $z_j^{\mathbf{k}} = \mathcal{O}(\delta^{e(l)+1-2\nu})$ for $\mathbf{k} \neq \pm\langle j \rangle$, since $z_{j,e(l)}^{\mathbf{k}} = 0$. For $\mathbf{k} \neq \pm\langle l \rangle$ this follows from (28) and (35) since $\mu(\mathbf{k}) > e(l)$, and for $\mathbf{k} = \pm\langle l \rangle$ from (28) and (38),
- $z_l^{\pm\langle l \rangle} = \mathcal{O}(\delta^{e(l)})$ by (41) and $\dot{z}_l^{\pm\langle l \rangle} = \mathcal{O}(\delta^{e(l)+1-2\nu})$ by (28) and (35).

This shows that this difference is in fact of order $\mathcal{O}(\delta^{2e(l)+1-2\nu})$. The improved order for $l = 1$ follows from the fact that in this case $\dot{z}_{1,2}^{\pm\langle 1 \rangle} = 0$ (see Example 9), and hence $\dot{z}_1^{\pm\langle 1 \rangle} = \mathcal{O}(\delta^{e(1)+2(1-2\nu)})$. To get the summed estimate, we proceed as in the proof of Theorem 4 in Section 4.4: We multiply the difference with $\sigma_l \delta^{-2e(l)}$ and sum over l , we use (42), and we apply the Cauchy–Schwarz inequality and the estimates of Lemma 10. \square

Combining Lemmas 15–17 yields the statement of Theorem 7, provided that $\delta^{1-2\nu}$ is sufficiently small.

Acknowledgement

We gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft (DFG) through CRC 1173, CRC 1114 and project GA 2073/2-1.

References

- [1] W. BAO, X. DONG, [Analysis and comparison of numerical methods for the Klein-Gordon equation in the nonrelativistic limit regime](#), *Numer. Math.* **120** (2012), 189–229.
- [2] B. CANO, [Conservation of invariants by symmetric multistep cosine methods for second-order partial differential equations](#), *BIT* **53** (2013), 29–56.
- [3] B. CANO, [Conserved quantities of some Hamiltonian wave equations after full discretization](#), *Numer. Math.* **103** (2006), 197–223.
- [4] B. CANO, A. GONZÁLEZ-PACHÓN, [Plane waves numerical stability of some explicit exponential methods for cubic Schrödinger equation](#), *J. Comp. Math.* **34** (2016), 385–406.
- [5] D. COHEN, L. GAUCKLER, E. HAIRER, C. LUBICH, [Long-term analysis of numerical integrators for oscillatory Hamiltonian systems under minimal non-resonance conditions](#), *BIT* **55** (2015), 705–732.
- [6] D. COHEN, E. HAIRER, C. LUBICH, [Conservation of energy, momentum and actions in numerical discretizations of non-linear wave equations](#), *Numer. Math.* **110** (2008), 113–143.
- [7] M. DAHLBY, B. OWREN, [Plane wave stability of some conservative schemes for the cubic Schrödinger equation](#), *M2AN Math. Model. Numer. Anal.* **43** (2009), 677–687.
- [8] P. DEUFLHARD, [A study of extrapolation methods based on multistep schemes without parasitic solutions](#), *Z. Angew. Math. Phys.* **30** (1979), 177–189.
- [9] X. DONG, [Stability and convergence of trigonometric integrator pseudospectral discretization for \$N\$ -coupled nonlinear Klein–Gordon equations](#), *Appl. Math. Comput.* **232** (2014), 752–765.
- [10] E. FAOU, L. GAUCKLER, C. LUBICH, [Plane wave stability of the split-step Fourier method for the nonlinear Schrödinger equation](#), *Forum Math. Sigma* **2** (2014), e5, 45 pp.
- [11] E. FAOU, B. GRÉBERT, [Hamiltonian interpolation of splitting approximations for non-linear PDEs](#), *Found. Comput. Math.* **11** (2011), 381–415.
- [12] E. FAOU, B. GRÉBERT, E. PATUREL, [Birkhoff normal form for splitting methods applied to semilinear Hamiltonian PDEs. I. Finite-dimensional discretization](#), *Numer. Math.* **114** (2010), 429–458.

- [13] B. GARCÍA-ARCHILLA, J. M. SANZ-SERNA, R. D. SKEEL, [Long-time-step methods for oscillatory differential equations](#), *SIAM J. Sci. Comput.* **20** (1999), 930–963.
- [14] L. GAUCKLER, [Error analysis of trigonometric integrators for semilinear wave equations](#), *SIAM J. Numer. Anal.* **53** (2015), 1082–1106.
- [15] L. GAUCKLER, E. HAIRER, C. LUBICH, D. WEISS, [Metastable energy strata in weakly nonlinear wave equations](#), *Comm. Partial Differential Equations* **37** (2012), 1391–1413.
- [16] H. GRUBMÜLLER, H. HELLER, A. WINDEMUTH, K. SCHULTEN, [Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions](#), *Mol. Sim.* **6** (1991), 121–142.
- [17] E. HAIRER, C. LUBICH, [Spectral semi-discretisations of weakly nonlinear wave equations over long times](#), *Found. Comput. Math.* **8** (2008), 319–334.
- [18] E. HAIRER, C. LUBICH, G. WANNER, [Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations](#), vol. 31 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.
- [19] M. KHANAMIRYAN, O. NEVANLINNA, T. VESANEN, [Long-term behavior of the numerical solution of the cubic non-linear Schrödinger equation using Strang splitting method](#), preprint (2012), <http://www.damtp.cam.ac.uk/user/na/people/Marianna/papers/NLS.pdf>.
- [20] T. I. LAKOBA, [Instability of the split-step method for a signal with nonzero central frequency](#), *J. Opt. Soc. Am. B* **30** (2013), 3260–3271.
- [21] M. TUCKERMAN, B. J. BERNE, G. J. MARTYNA, [Reversible multiple time scale molecular dynamics](#), *J. Chem. Phys.* **97** (1992), 1990–2001.
- [22] J. A. C. WEIDEMAN, B. M. HERBST, [Split-step methods for the solution of the nonlinear Schrödinger equation](#), *SIAM J. Numer. Anal.* **23** (1986), 485–507.